DETERMINING AND CHARACTERIZING IMMUNOLOGICAL SELF/NON-SELF

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Ying Li

©Ying Li, February 2007. All rights reserved.

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

ABSTRACT

The immune system has the ability to discriminate self from non-self proteins and also make appropriate immune responses to pathogens. A fundamental problem is to understand the genomic differences and similarities among the sets of self peptides and non-self peptides. The sequencing of human, mouse and numerous pathogen genomes and cataloging of their respective proteomes allows host self and non-self peptides to be identified. T-cells make this determination at the peptide level based on peptides displayed by MHC molecules.

In this project, peptides of specific lengths (k-mers) are generated from each protein in the proteomes of various model organisms. The set of unique k-mers for each species is stored in a library and defines its "immunological self". Using the libraries, organisms can be compared to determine the levels of peptide overlap. The observed levels of overlap can also be compared with levels which can be expected "at random" and statistical conclusions drawn.

A problem with this procedure is that sequence information in public protein databases (Swiss-PROT, UniProt, PIR) often contains ambiguities. Three strategies for dealing with such ambiguities have been explored in earlier work and the strategy of removing ambiguous k-mers is used here.

Peptide fragments (k-mers) which elicit immune responses are often localized within the sequences of proteins from pathogens. These regions are known as "immunodominants" (i.e., hot spots) and are important in immunological work. After investigating the peptide universes and their overlaps, the question of whether known regions of immunological significance (e.g., epitope) come from regions of low host-similarity is explored. The known regions of epitopes are compared with the regions of low host-similarity (i.e., non-overlaps) between HIV-1 and human proteomes at the 7-mer level. Results show that the correlation between these two regions is not statistically significant. In addition, pairs involving human and human viruses are explored. For these pairs, one graph for each k-mer level is generated showing the actual numbers of matches between organisms versus the expected numbers. From graphs for 5-mer and 6-mer level, we can see that the number of overlapping occurrences increases as the size of the viral proteome increases.

A detailed investigation of the overlaps/non-overlaps between viral proteome and human proteome reveals that the distribution of the locations of these overlaps/non-overlaps may have "structure" (e.g. locality clustering). Thus, another question that is explored is whether the locality clustering is statistically significant. A chi-square analysis is used to analyze the locality clustering. Results show that the locality clusterings for HIV-1, HIV-2 and Influenza A virus at the 5-mer, 6-mer and 7-mer levels are statistically significant. Also, for self-similarity of human protein Desmoglein 3 to the remaining human proteome, it shows that the locality clustering is not statistically significant at the 5-mer level while it is at the 6-mer and 7-mer levels.

ACKNOWLEDGEMENTS

With respect and gratitude, I acknowledge my supervisor Dr. Anthony Kusalik, whose expertise, guidance, support, encouragement, and patience has made this dissertation possible. His feedback on my thesis research has vastly improved the quality of this dissertation and provided an enthralling educational experience. I would also like to thank Dr. Mik Bickis from Department of Mathematics and Statistics providing support on statistical analysis. A very special thank goes out to Dr. Shawn Babiuk. He helped me with immunological background knowledge and the outline of my thesis from an immunological perspective.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

APC	Antigen Presenting Cell
BCR	B Cell antigen Receptor
DC	Dendritic cell
ER	Endoplasmic Reticulum
MHC	Major Histocompatibility Complex
TAP	Transport Associated Protein
TCR	T Cell antigen Receptor

CHAPTER 1

INTRODUCTION

Immuno-informatics is an area where people use bioinformatics tools to understand the immune system better as well as to improve the development of immunotherapies. Just as DNA is the information system for life, peptides are the information system for the adaptive immune system. The immune system has multiple layers of defense to protect against pathogens. The initiation, regulation and termination of an immune response involves a large number of cells of different types and several stimulatory/inhibitory signals delivered locally and systemically [4, 22].

The immune system has the ability to discriminate self from non-self proteins and then make appropriate immune responses to pathogens. Such an ability is an important property in maintaining tissue/organism integrity. Breakage this self-tolerance is one of the main bases for autoimmune diseases [10, 13, 21]. Self proteins are proteins from the organism itself while non-self proteins are proteins not from the organism but from such sources as invading pathogens. According to this characteristic of proteins, the immune system decides whether or not to respond to infections, and which type of response to make. An infected cell can "present" peptides that are generated from the degradation of non-self proteins to immune cells in the organism. The presentation of peptides to the T-cells is done by MHC molecules, which have one of the largest degrees of polymorphisms among mammalian proteins. That is, within the "groove" of the MHC molecules where the peptide fragments bind, there may be various shapes in order to bind peptides.

Vertebrate immune systems process self and non-self proteins into peptide fragments 8 - 25 amino acids long, which are presented to the T-cell repertoire by surface MHC molecules [7, 9]. There are different types of MHC molecules such as MHC class I and MHC class II molecules. They have different conformational shapes, and different binding preferences. The typical length of peptides presented to $CD8^+$ T cells by MHC class I molecules is 9 residues; the peptides presented to $CD4^+$ T cells by MHC class II molecules tend to be longer, with a typical length of 12 - 20 residues. In addition, the typical length of peptides presented to B cells is 5 - 6 residues. Thus, in this work, self and non-self proteins are computationally fragmented into sets of peptides of specific lengths (k-mers).

In this research, "host proteome" refers to the human proteome. Host self consists of all the possible peptides that can be generated from the host proteome, while non-self consists of all the possible peptides that can be generated from the proteome of a foreign organism. Overlaps refer to the occurrence of identical peptides within the proteomes of different organisms [3].

The MHC-I and MHC-II as well as B cell immune response mechanisms are important to this work. Chapter 2 therefore presents a review of the immune system and discusses relevant details of the T-cell and B-cell immune response mechanisms. The function of the immune system as well as its components are discussed.

Having the knowledge of immune response mechanisms, we may consider the question of how to use computers and information techniques to help to understand the role of the immune system in infectious diseases, autoimmune diseases and cancers. Therefore, in the following chapters, such techniques are proposed, investigated, or compared. Two topic problems are proposed: one is to investigate the correlation between regions of immunological significance and regions of low host-similarity; the other is to explore the statistical significance of locality clustering of the overlapping occurrences within the viral proteome.

In Chapter 3, we introduce the idea of "peptide universe" and its role in immunoinformatics study. The chapter goes on to present how the peptide universes for various species are generated, including detail about the data structures and algorithms involved. Finally, the chapter discusses how the overlaps between proteomes of pathogen and host is determined. These overlaps include unique ones as well as occurrences involving duplicates.

Unfortunately, the protein sequences in protein databases are not perfect. Sometimes there are ambiguities as to which amino acid occurs at a particular position. Presence of ambiguities can affect the results of comparisons of self, non-pathogenic non-self and pathogenic non-self peptide universes. In Chapter 4, three types of filtering strategies for dealing with such ambiguities are introduced and compared based on time efficiency at the 9-mer level. For each strategy, advantages and disadvantages as well as potential problems are discussed. The strategy of removing only generated k-mers which contain ambiguous amino acids (B, X or Z) is used in subsequent stages of our work.

Within the sequences of antigens, there exist regions of peptide fragments which elicit immune responses. These regions are known as "immunodominants" and are important in immunological work. In chapter 5, the question of whether known regions of immunological significance (e.g. epitopes, agretopes) in pathogenic proteomes come from regions of low host-similarity is explored. If the answer is "yes", it means that the levels of peptide overlap may be used to predict possible regions of immunological significance. The known regions of epitopes are compared with the regions of low host-similarity (i.e., non-overlaps) between HIV-1 and human proteomes at the 7-mer level. Results show that the correlation between these two regions is not statistically significant. In addition, pairs of organisms involving human and human viruses are explored. For these pairs, one graph for each k-mer level is generated showing the actual numbers of matches between organisms versus the expected numbers. At random, one would expect an approximate linear relationship between the size of the viral proteome (m) and the number of viral k-mers including repeats occurring in the human proteome providing that $m \ll 20^k$. From graphs for 5-mer and 6-mer level, we can see that the number of overlapping occurrences increases as the size of the viral proteome increases. Such an observation motivates further bench and computational investigation.

In chapter 6, a detailed investigation of the overlaps/non-overlaps between viral proteome and human proteome reveals that the distribution of the locations of these overlaps/non-overlaps has "structure". Here "structure" refers to locality clustering. Thus, another question that is investigated is whether the locality clustering is statistically significant. A chi-square analysis is used to analyze the locality clustering. Results show that there exists structure for HIV-1, HIV-2 and Influenza A virus at the 5-mer, 6-mer and 7-mer levels. Also, for self-similarity of human protein Desmoglein 3 to the remaining human proteome, it shows that there is structure at the 7-mer level.

Extended research and future work are discussed in chapter 7. For instance, the degrees of overlap between the generated peptide universes could be used as for building phylogenetic trees especially relevant to the immunological context.

The complete list of all the proteomes and raw data that are used in the research work are listed in Appendices.

CHAPTER 2

OVERVIEW OF IMMUNE SYSTEM

This chapter is a general introduction of the immune system. Readers who are familiar with the immune system may skip this chapter. Those who would like further information are referred to texts by Janeway et al. [11], Coico et al. [4], or Thomas et al. [12].

2.1 Immune System

The major assignment of an immune system is to defend against infections. The evolution of pathogens induces strong selection pressure on host immune systems. The most advanced result of this co-evolution is found in higher vertebrates. Vertebrate immune systems provide rapid, specific, protective immune responses to infectious bodies without causing damage to the hosts themselves. In addition, these immune systems can "remember" a pathogen and induce a protective response in the event of subsequent exposure.

Vertebrate immune systems have two branches: innate and adaptive immunity. The former is phylogenetically older and exists in a primitive form in all multicellular organisms. The later is about 400 million years old and is found in cartilaginous and bony fish, amphibians, reptiles, birds and mammals [26]. The innate immune system distinguishes between self and non-self according to complexes such as carbohydrate signals [8]. Compared to this relatively non-specific approach, adaptive immunity, which is induced by lymphocytes, generates a very large repertoire of antigen receptors (either T Cell Antigen Receptor (TCR) or B Cell Antigen Receptor (BCR)) with the potential to recognize different antigens. Adaptive immunity can be further divided into two types. One type is called humoral immunity, which is mediated by antibody molecules secreted by B lymphocytes. The other type is called cellular immunity, which is mediated by T lymphocytes. An essential difference between these two types of immunity is the means by which they recognize pathogens.

Adaptive immunity is an acquired and highly specific immunity. Its main characteristic is the use of antigen-specific receptors on T and B cells to drive targeted, two-stage effector responses. As shown in Figure 2.1, in the first stage, the antigen is presented to and recognized by the antigen specific T or B cell leading to cell priming, activation and differentiation. These processes usually occur within the specialized environment of lymphoid tissue. In the second stage, the effector response takes place, either due to the activated T cells leaving the lymphoid tissue and going to the disease site, or due to the release of antibodies

from activated B cells into blood and tissue fluids, and hence to the infective part. The specificity of the antigen receptors can be predicted from the amino acid sequence. This specificity can be used to select epitopes in making some vaccines. Portions from antigen molecules with which an antibody or lymphocyte react are called epitopes. T cells show high antigen specificity, but also a vast receptor diversity. The antigen specificity in T cells is similar to B cells. Finally, the T cells only recognize antigens that are presented on other cells in association with MHC molecules, the receptor should be a membrane-bound molecule.

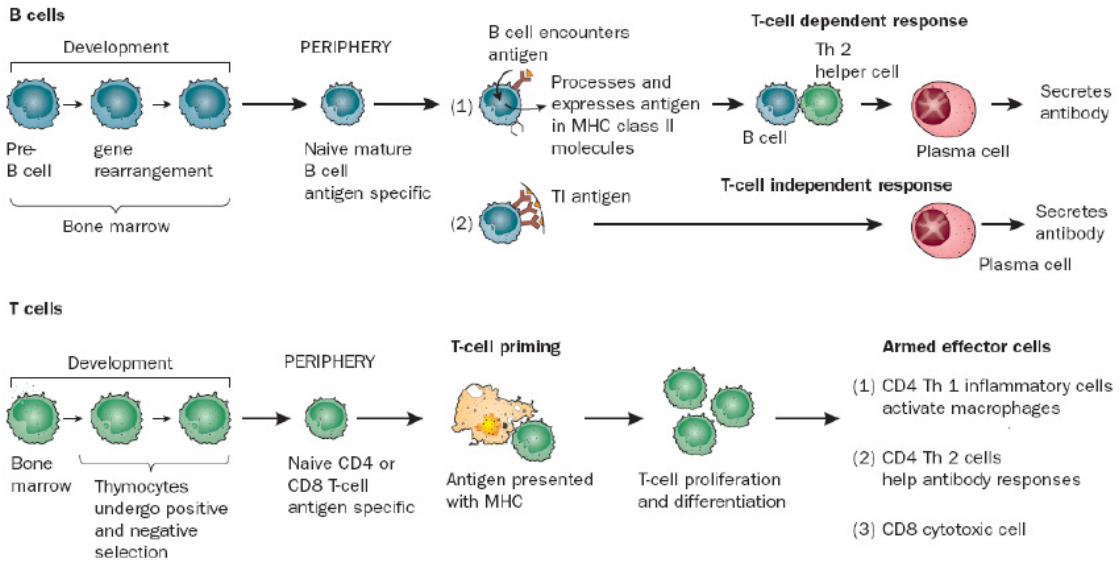

Figure 2.1: The role of T and B lymphocytes in specific immunity. The figure is taken from Rammensee et al. [22].

All the components of an immune system work together in order to make efficient and effective defenses. For instance, the innate immunity may instruct the adaptive immunity system regarding that to what component it responds [8]. Thus, the decisions of whether and how to defend are influenced by multiple factors and each of them are controlled by one or more parts of the host immune system.

In addition to defense, another important assignment of vertebrate immune system is tolerance to host self and homeostasis. Host self consists of all the possible peptides that can be generated from the host proteome. Even though self-reactive lymphocytes are created constantly, they are removed within the thymus gland so as to prevent auto-immune disease. Homeostasis is the ability or tendency of an organism or a cell to maintain internal equilibrium by adjusting its physiological processes. The immune system maintains such an equilibrium state where the number of immune cells is roughly the same as it was before the defense, although it is continuously being exposed to self antigens and generating responses to a diverse collection of microbes. To maintain this homeostasis, the repertoire of immune cells is altered in a way that ensures a protective response to a particular antigen.

2.2 MHC Class I and MHC Class II Antigen Processing and Presentation Pathways

As shown in Figure 2.2, there are two main pathways to process and present antigens to lymphocytes. The MHC class I pathway, or the endogenous pathway, stimulates cell-mediated immunity. In order to present endogenous antigens to $CD8^+$ T cells, a precursor peptide must be generated by some proteasome. This peptide may be trimmed at the N-terminal by other peptidases in the cytosol [24]. It must then bind to a transporter molecule called "transporter associated with antigen processing (TAP)" in order to be translocated to the endoplasmic reticulum (ER). Here its N-terminal can be trimmed by the amino-peptidase associated with antigen processing. Then it binds to an MHC I molecule and the complex is transported to the cell surface [25]. Once there, this complex is recognized by MHC class I restricted $CD8^+$ cytotoxic T cells which in turn are activated, amplified and attack cells containing the antigen. This allows for a cell-mediated immune response to "see" foreign proteins within all nucleated cells. In this way, cells infected with a virus may be detected and destroyed. Selectivity is exercised at all of the previous steps. For instance, only about half of the peptides that are presented on the cell surface by MHC molecules are recognized by TCRs. The most selective step is binding to the MHC I molecule, wherein only 0.5% of the total 20^k peptides bind with an affinity strong enough to generate an immune response [29]. Therefore, in order for a peptide to be immunogenic (immunological significant), it must be "special" as compared to other peptides produced in a given cell.

The presentation on Class II MHC molecules follows a different pathway [2]. It is also called the antibody pathway or the exogenous pathway. This pathway processes exogenous antigens that are taken up by cells and presented in such a way as to stimulate helper T cells, which then stimulate B lymphocytes and antibody production (the humoral pathway). Precursor MHC class II molecules accumulate in endosomal compartments in the Golgi complex. Here one chain of the molecules is degraded, leaving the MHC-II molecules free to bind peptides derived from endocytosed antigens. The peptide class II complexes are subsequently transported to the cell surface for presentation to $CD4^+$ T cells. $CD4^+$ activation leads to production of cytokines which in turn activate a wide range of cells around them. The reaction therefore needs to be kept in check, which is achieved by only a small number of class II antigen-presenting cells being able to drive the response. MHC-II molecules are much like antibodies in that they bind to antigens. The exogenous pathway is well-suited for detecting bacterial infections, which are primarily extracellular. Viral infections are primarily intra-cellular and more easily detected by the MHC I pathway.

2.3 Length of Peptides Recognized by Lymphocytes

Both MHC I and MHC II are highly polymorphic. They have different conformational shapes and binding preferences. The typical length of peptides presented to $CD8^+$ T cells by MHC class I molecules

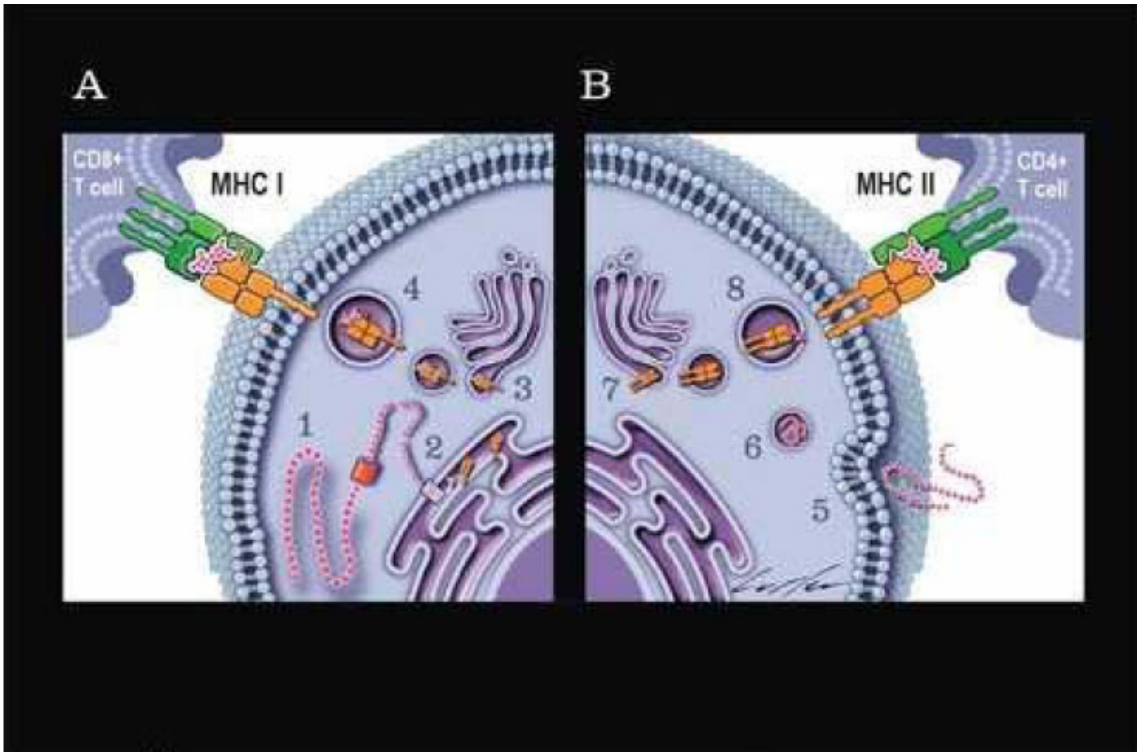

Figure 2.2: I. Presentation of antigens to the immune system. The MHC class I system mediates presentation of endogenous or intra-cellularly-produced proteins. The MHC class II system mediates presentation of exogenous or externally-produced antigens. The MHC I pathway is used predominantly for immune presentation of endogenously synthesized proteins. Intra-cellular or endogenously produced proteins (1) are broken down into peptides (8 to 13 amino acids long) by spliceosomes and then directed into the endoplasmic reticulum (ER) through transport associated protein (TAP) molecules (2). Once in the ER, the peptide antigens bind to activated MHC class I molecules (3), which subsequently transport the complexes through the Golgi complex toward the cell surface (4). This pathway generates an immune response biased toward cell-mediated/cytotoxic $CD8^+$ lymphocyte activation. **II**. Following endocytosis, the MHC II pathway presents peptide antigens which originated outside of the cell. Extracellular or exogenous antigens are taken up into specialized antigen presenting cells (APCs) by endocytosis (5). The antigen is degraded in a lysosome (6) into immunogenic peptides which bind to MHC class II molecules (7). To produce an antibody biased immune response, the MHC class II-peptide antigen complex is transported to the cell surface (8) where it binds to $CD4^+$ helper T cells. The figure is taken from Rammensee et al. [22].

is 9 residues; the peptides presented to $CD4^+$ T cells by MHC class II molecules tend to be longer, with a typical length of 12 - 20 residues. For class I MHC molecules, at least two of the nine residues are known as "anchor residues" and have decisive effects in the binding process. Once the peptides have bound to the corresponding positions in the groove of the MHC class I molecules, the peptide-MHC complex will be transported to the surface of the cell and presented to $CD8^+$ T cells. The T cells then recognize the peptide-MHC complex and initiate cell destruction. Each MHC class I molecule is able to bind various peptides, but will bind certain peptides with greater affinity. The binding mechanisms of MHC class II molecules are similar to those of MHC class I molecules, but with longer peptides and the fragments are normally derived from endocytosed antigens.

The B cell receptors recognize antigen differently than T cell receptors. The antibody recognizes the conformational structure (shape) of epitopes, and such antigens do not require processing. The typical length of peptides presented to B cells is 5 - 6 residues.

CHAPTER 3

PEPTIDE UNIVERSES AND CORRESPONDING OVERLAPS

This chapter describes the data which is used in the study, as well as the methodology by which it is derived. The generation of a peptide universe is presented in Section 3.1. Overlap data, which is the set of k-mers that occur in two proteomes, is described in Section 3.2 and its derivation is discussed in Section 3.3. In Section 3.4, the research goals to be answered using the above data are stated.

3.1 Peptide Universe Generation

The various immunological pathways operate on peptide fragments of different lengths. For example, MHC I molecules bind peptides ranging from 8 to 13 amino acids long, with preference for those of length 9. On the other hand, 5 - 6 amino acids are the minimum requisite to induce an antibody response in B cells. To study the peptides that are involved in one of these pathways, it is necessary to derive the sets of peptides of specific lengths. A peptide universe is the set of all possible unique peptide fragments of a specific length k derived from an organism's proteome. Each proteome of an organism will have multiple peptide universes, one for each length k.

All the steps in generating a peptide universe are shown in Figure 3.1. A "FASTA" file (i.e., a file whose name ends in ".FASTA") contains the protein sequences of an organism. From it, a list of k-mers is generated for every protein (of length n) by choosing each position (except the last $n-1$) in the sequence as the starting point. The resulting set of k-mers is stored in an "nmerext" file (name ends in ".nmerext"), which contains all the k-mers that occurred at consecutive locations in all sequences. Specific k-mers may occur multiple times, and such repeated k-mers are included in the "nmerext" file. Based on the latter file, an "intraseq_repeats" file (name ends in ".intraseq_repeats") and an "interseq_repeats" file (name ends in ".interseq_repeats") are produced. These files contain all the k-mers that are repeated within a protein sequence, and all the k-mers that are repeated in two or more protein sequences, respectively. Combining the above two sets of k-mers yields an "all_repeats" file (name ends in ".all_repeats"), which contains all k-mers that are repeated in the k-mer universe of an organism, as either an inter-sequence or intra-sequence repeats. Finally, by including a unique instance of all k-mers from the "nmerext" file, a "library" file (name ends in ".library") is generated. This "library" file defines the peptide universe (for a given k) of a given species in our study. The "library" file contains all the unique k-mers in the proteome of an organism.

3.2 Overlaps Between Two Proteomes

After the peptide universe for each organism is obtained, the overlaps and non-overlaps between two proteomes can be determined. The flow chart in Figure 3.2 shows the process in detail. The method is to use the "library" file of one species (a list of all k-mers in its peptide universe) to scan through the proteome of the other species (stored in a "FASTA" file) looking for k-mers which occur in both. These k-mers are called overlaps or matches. Here the species whose "library" file is used is known as "BUG" and the species whose "FASTA" file is used is known as "HOST". For each k-mer in the peptide universe of "BUG", an indication of whether that k-mer is present in the proteome of "HOST" is recorded in a "BUG_vrs_HOST.peptsrch" file (i.e., a file whose name ends with ".peptsrch"). Each time the peptide is found in "HOST", the protein in which the instance is found and the position of the occurrence in that protein are recorded. Here exact matches are sought, not "similar" or "approximate" ones as might be done with sequence alignment. Based on the search for peptides, the overlap/intersection between "HOST" and "BUG" is produced and all the k-mers that occurred in both organisms are included in a "BUG_inters_HOST" file (name follows the pattern "BUG_inters_HOST"). On the other hand, those k-mers that exist in the peptide universe of "BUG" but do not occur in the proteome of "HOST" are included in a "BUG_diff_HOST" file (name follows the pattern "BUG_diff_HOST").

Two different techniques are used to count the number of k-mers in the overlap of both proteomes. One technique counts the *distinct* k-mers in the overlap of two proteomes. The other technique counts the number of k-mers in the overlap (including *duplicates*) of both proteomes. The first technique only considers distinct k-mer (from BUG's peptide universe) occurrences in HOST's proteome. That is, only one occurrence of a k-mer is counted even if it appears several times in HOST's proteome file. Alternatively, the number of k-mers in the overlap (including duplicates) of both proteomes can be counted. For example, if a k-mer from BUG's peptide universe occurred 5 times in HOST's proteome, the number of occurrences will be 5. The first technique may cause some problems. For instance, if a k-mer appears 100 times, which means it is very important and has quite high frequency, it is still counted only once. That is, each k-mer in the library of one proteome is all treated the same. However, measures of overlaps should capture the situation when one k-mer appears multiple times. Therefore, the repeating of a k-mer is considered in the second technique. However, the second technique does not reflect the frequency distribution for each k-mer from BUG's peptide universe, only the total number of occurrences. For instance, a certain k-mer may appear 100 times either only in one protein of HOST or randomly distributed through the HOST's proteome. Thus both methods of counting overlaps have shortcomings.

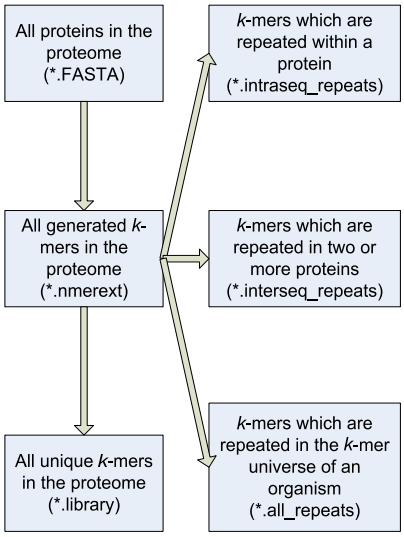

Figure 3.1: Data flow chart of peptide universe generation for a given species.

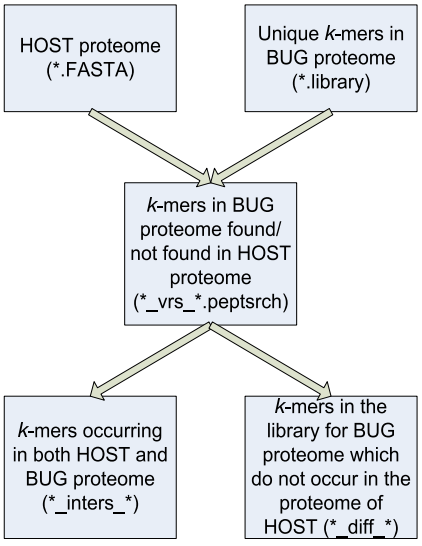

Figure 3.2: Data flow chart for the generation of overlaps and non-overlaps between two proteomes, i.e. those for organisms “HOST” and “BUG”.

3.3 Computational and Algorithmic Issues

Since MHC-I molecules bind peptides ranging from 8 - 10 amino acids long and there are 20 different amino acids, the size of the peptide universes vary from 20^8 to 20^{10} (2.56×10^{10} to 1.024×10^{13}) peptides. There are 5.12×10^{11} potential peptides using the 20 standard amino acids which represent the 9-mer peptide universe. A similar result holds for peptides 5 to 6 amino acids long in the pathway of B cells. There are 3.2×10^6 to 6.4×10^7 potential peptides representing the peptide universe when using the standard 20 amino acids. The longer the length of an k-mer peptide, the larger the size of the corresponding peptide universe.

When dealing with data of large size, time efficiency becomes an issue. When searching a peptide from BUG's peptide universe in HOST's proteome, if a linear search algorithm is used for a given proteome BUG of length m and proteome HOST of length n, time O($m \times n$) is required. With such an algorithm, it will take a fairly long time to generate overlaps between two proteomes with data of large size (i.e., mouse and human). Therefore, a data structure called a "suffix tree" [6] is employed in order to make the search more efficient when using the peptide universe of BUG to scan through the proteome of HOST.

In general, any string of length l can be degenerated into l suffixes, and these suffixes can be stored in a suffix tree. Creating this structure uses time O(l^2) and searching for a pattern of length p in it requires time O(p). For example, consider the suffix tree corresponding to the string "ATCATG". It generates the following suffixes: G, TG, ATG, CATG, TCATG and ATCATG. Then the suffix tree in Figure 3.3 is produced accordingly. Suppose we search for pattern "TG", it will take two steps to find the node "TG" [6]. Moreover, a more sophisticated algorithm can be used to build the suffix tree in time O(l).

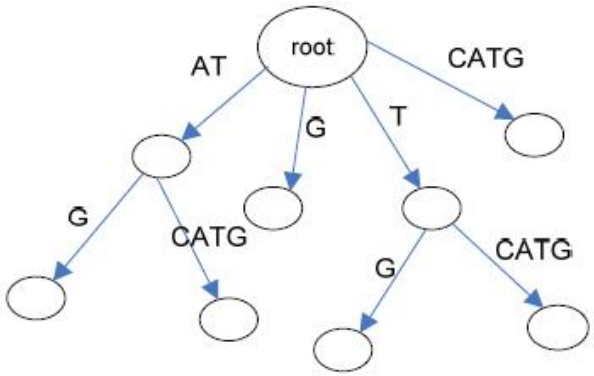

Figure 3.3: An example of suffix tree that are constructed for string "ATCATG".

In our work, the entire proteome for HOST (length n) is stored in a suffix tree and the implementation of building the suffix tree uses time O(n). It is built once for each run of the program from the proteome of HOST and can be used for as many queries as needed. In this case each k-mer in the peptide universe of BUG is used to search for the same k-mer in the suffix tree. Only k comparisons are needed to determine whether the k-mer is found or not, which is extremely time efficient. Thus, this suffix tree-based search algorithm requires O($k \times S_B$), where k is the length of the peptide fragment in query and S_B is the size

of the peptide universe for proteome BUG. That is, after peptide universe of proteome BUG (length m) is generated, there are S_B queries and each query takes O(k) time. Since k is a constant from 5 to 9, the suffix tree-based search actually requires only a linear time O(S_B). These efficient programs were developed by Chris Lewis of the Bioinformatics Research Lab at the University of Saskatchewan (for details see http://homepage.usask.ca/∼ctl271/857/suffix_tree.shtml).

3.4 Discussion

Size of overlaps between a host's (usually human) and a pathogen's 9-mer peptide universe varies from 0.1% - 0.2% with most pathogens, compared to the size of the peptide universe of the human proteome. While size of overlaps between hosts (e.g. human and mouse), as well as between different bacterial species is much higher, ranging from 30% - 40% [3]. These results are consistent with the idea that the overlap between a host's and pathogen's peptide universe tends to be small, whereas overlap between peptide universes of mammalian species or similar pathogens tends to be much larger due to the ability of immune system to descriminate self from nonself proteins. Overlaps of peptides from different organisms can have important immunological consequences [3]. If a foreign peptide is also a self peptide, either the overall T-cell pathway is tolerant to it or any T cell responses to it may lead to an autoimmune response. By determining a host's self universe and incorporating an epitope prediction model, we may form the basis of a rapid, inexpensive and computationally driven system for the individuation of antigenic sequences that are the targets of autoimmune responses [16]. Consequently, this proteomic strategy may serve as a general method suitable to define/distinguish/screen disease-relevant or disease-irrelevant epitopes within potential antigens [15]. Therefore, the level of potentially cross-reactive k-mers between pathogen and host is worth investigating.

The k-mer peptides in the overlap between different pathogens are potentially cross-protective antigens. If an immune response is generated in the host to an k-mer of one pathogen, the host will generate a secondary immune response against the identical k-mer from another pathogen. Comparing the overlap of k-mers between different pathogens demonstrates the level of potential cross-protective antigens and is also valuable to investigate.

The peptides in the overlap between host and pathogen peptide universes are not useful as vaccine candidates since (1) immune responses are difficult to generate for these epitopes due to deletion of specific T cells and (2) if immune responses are generated to these epitopes, self-reactive cells will be generated leading to autoimmunity. However, the information is still useful for vaccine design because it indicates peptides which should be avoided as vaccine candidates.

CHAPTER 4

COMPARISON OF DIFFERENT FILTERING STRATEGIES

The protein sequences in protein databases are not perfect. Sometimes there are ambiguities during sequencing as to which amino acid occurs at a particular position. These ambiguities are coded by characters similar to those used to code unambiguous amino acids. Three types of ambiguous amino acids are considered in the literature and the characters used to code them are listed in Table 4.1.

Table 4.1: Ambiguous amino acids and their corresponding representational characters

Character	**Ambiguous Amino Acids**
B	Asparagine or Aspartic acid
Z	Glutamine or Glutamic acid
X	Unknown Amino Acid

Presence of ambiguities can adversely affect the results of the comparisons of self, non-pathogenic non-self and pathogenic non-self peptide universes. An ambiguity is considered as a unique symbol when comparing k-mers; i.e. two k-mers containing ambiguities are considered to match if all symbols are the same in all positions, including any instances of B, X, or Z. For instance, "CLXMHBZKC" and "CLXMHBZKC" are considered as match while 'X', 'B' and 'Z' in these two 9-mers could represent different amino acids. For 'B' and 'Z', there is a 50% probability that this match is correct while for 'X', there is only a 5% possibility that this match is correct. Thus, the ambiguities need to be filtered from proteome information used to determine peptide universes or proteome overlap. Three types of filtering strategy are used in this work to process all the k-mers of a given proteome. Comparison of these three filtering methodologies are only done for 9-mers, and the results are expected to be similar when we extend to other values of k.

4.1 Three Filtering Strategies

Within the overall goal of characterizing self versus non-self peptides, we need to address the presence of ambiguous amino acids in the sequence data. The focus here is to compare the merits of three possible methods to deal with the ambiguity. Three filtering strategies are examined in the research work as follows:

- The first is to include all proteins, irrespective of whether they contain ambiguous amino acids (B, X or Z). Note that this would be considered ignoring the ambiguities.

- The second is to remove any proteins that contain any ambiguous amino acids (B, X or Z) [3]. Note that no k-mers would be generated from such a protein during the analysis.

- The third is to remove only generated k-mers which contain ambiguous amino acids (B, X or Z).

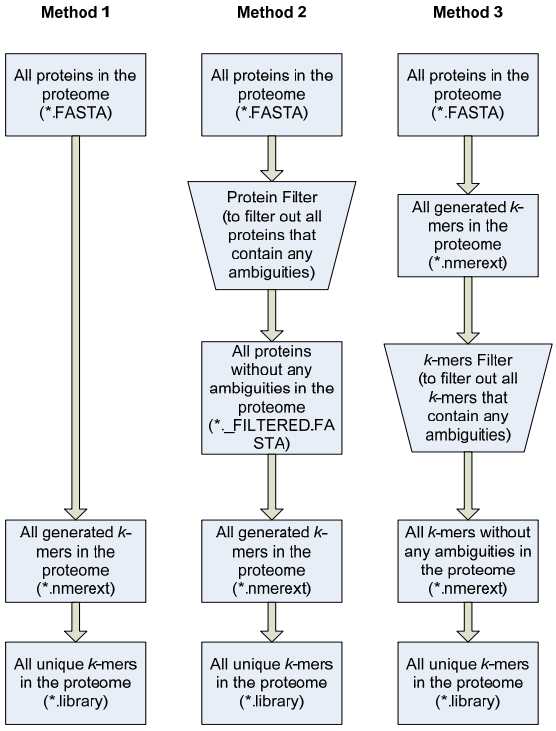

Figure 4.1: Data flow chart of three filtering strategies in the generation of peptide universes for a given species. Method 1 does not perform any filtering; Method 2 includes a protein filter to remove any proteins that contain ambiguous amino acids; Method 3 includes an k-mer filter to remove generated k-mers which contain ambiguous amino acids.

As shown in Figure 4.1, the first approach does not use any filtering. In the second approach, a protein filter is added before the generation of the peptide universe. That is, given a "FASTA" file, a "filtered FASTA" file (i.e., a file whose name ends in "_FILTERED.FASTA") containing all the protein sequences without any ambiguities is generated and then used as the input file to produce the peptide universe. In the third approach, a k-mer filter is added after the generation of all the k-mers of a given proteome. That is, after generating all the k-mers, we remove the k-mers which contain ambiguous amino acids.

The first approach which does not do any filtering causes errors of mismatch as well as errors of missing information. For example, suppose one peptide from BUG is "CLXMHNLKC" and two peptides from HOST are "CLXMHNLKC" and "CLAMHNLKC". As we know, "X" is considered as unknown amino acid and thus it can actually represent any one of the 20 standard amino acids. However, in the first method, "X" has become the 21st standard amino acid. Therefore, "CLXMHNLKC" and "CLXMHNLKC" will be considered to match while "CLXMHNLKC" and "CLAMHNLKC" will not. In fact, "CLXMHNLKC" and "CLXMHNLKC" should not match if first "X" and second "X" refer to different amino acids. Alternatively, "CLXMHNLKC" and "CLAMHNLKC" should match if "X" refers to "A". Matching "CLXMHNLKC" and "CLXMHNLKC" when one should not is considered an error of mismatch (or false positive) while not matching "CLXMHNLKC" with "CLAMHNLKC" when "X" is "A" is considered an error of missing information (or false negative). In the worst case, if we have "XXXXXXXXX" appearing in either the BUG or HOST proteome, only "XXXXXXXXX" is considered to match and the actual count of overlaps between BUG and HOST is inaccurate. However, this strategy does not require any computational resources to perform filtering.

The second approach is to remove any proteins that contain any ambiguous amino acids. Thus no k-mers are generated from such a protein during the analysis and no erroneous matches would be reported (e.g. the case "CLXMHNLKC" and "CLXMHNLKC" above). However, this filtering method causes errors of missing information. For organisms whose proteome consists of a single polyprotein, if the polyprotein contains even one ambiguous amino acid, the entire proteome is discarded. For organisms that have a limited number of proteins, removing proteins that contain ambiguities has a significant effect on any sequence-based analysis. In addition, this filtering method requires the cost of additional computational resources to do the filtering work though it takes less time to generate the overlaps.

The third approach is to remove generated k-mers with ambiguous amino acids. Intuitively, this eliminates errors of mismatch while retaining as much of the organism's proteome as possible. However, this filtering method also causes errors of missing information. In addition, it also requires the cost of additional computational resources to perform the filtering.

Now comes the question: which of the three approaches is the most desirable? On one hand, ambiguous amino acids should not be given biological meaning in the analysis. They can only make results more unpredictable and less accurate. In addition, since all the three filtering methods cause errors, the one with the least errors is more appropriate for our subsequent work. On the other hand, we may not want to spend significant computational resources on the filtering process. A decision on which strategy is more appropriate is therefore also dependent on the added computational resources necessary in each case. When determining which one is more appropriate, we consider the method with less errors first, and then check whether or not the additional computational cost is acceptable. The comparison of time efficiency among the three methods is given in Section 4.2.

During the comparison of the errors among the three filtering methods, we consider the errors of mismatch as false positives (FP) and the errors of missing information as false negatives (FN). First, both method 2 and method 3 don't have errors of mismatch, but method 2 has more errors of missing information than method 3. Therefore, method 3 is more appropriate than method 2. Then, comparing method 3 with method 1 reveals that method 1 has errors of mismatch while method 3 does not. Also, both methods have errors of missing information. Suppose m_H is the number of k-mers in HOST, m_H^* is the number of k-mers in HOST after filtering ambiguities, m_B is the number of k-mers in BUG and m_B^* is the number of k-mers in BUG after filtering ambiguities. Let Z be the total number of matches (including multiplicities) of matches in the filtered case. Then $\frac{Z}{m_H^* m_B^*}$ estimates the proportion of true matches between HOST and BUG proteomes. There are $(m_H - m_H^*)(m_B - m_B^*)$ potential matches left out. It is assumed that the proportion of potential matches has the same probability as the proportion of true matches. The number of missing matches is $(m_H - m_H^*)(m_B - m_B^*) \times \frac{Z}{m_H^* m_B^*}$ and the total number of true matches is $m_H m_B \times \frac{Z}{m_H^* m_B^*}$. Dividing the number of missing matches by the total number of true matches gives $(1 - \frac{m_H^*}{m_H})(1 - \frac{m_B^*}{m_B})$, which is the probability of errors of missing information in method 3. In method 1, the probability of errors of missing information is the probability ($P(ambiguity|match)$) of a k-mer with ambiguities given that this k-mer from BUG's proteome is a true match in the HOST's proteome. Then we have $P(ambiguity|match)$ = $1 - P(no\ ambiguity|match)$. It is assumed that the event of being a match between BUG and HOST as well as the event of k-mer containing ambiguity are independent. We also assume that the event of a k-mer from BUG's proteome containing an ambiguity and the event of a k-mer from HOST's proteome containing an ambiguity are independent. Therefore, $P(ambiguity|match)$ = $1 - P(no\ ambiguity)$ = $1 - P(no\ ambiguity\ in\ HOST)P(no\ ambiguity\ in\ BUG)$ = $1 - \frac{m_H^*}{m_H} \times \frac{m_B^*}{m_B}$. Subtracting the probability of errors of missing information in method 1 from method 3 gives

$$\frac{m_H^*(m_B^* - m_B) + m_B^*(m_H^* - m_H)}{m_H m_B}.$$

The result is negative and therefore method 1 has a higher probability of errors of missing information than method 3. Moreover, method 1 has errors of mismatch while method 3 does not. In conclusion, method 3 is the most appropriate filtering strategy among the three when considering the probability of errors.

In addition, ambiguities typically constitute a relatively small portion of a proteome. For instance, the percentage of ambiguous amino acids (including B, X, and Z) in *Homo sapiens* is only 0.01%. Other organisms contain fairly small numbers of ambiguities as well. Such information can be found at http://www.ebi.ac.uk/integr8 by examining the amino acid composition of a specific organism. What's more, although the second and third approaches result in information loss, this loss may have little affect on relative values since both HOST and BUG proteomes lose information.

4.2 Comparison of Three Strategies

4.2.1 Materials

A number of proteomes of model organisms were downloaded from the EBI website. All the bacterial and eukaryotic proteomes were from http://www.ebi.ac.uk/proteome, while viral proteomes were from http://www.ebi.ac.uk/genomes/virus.html. Among all these proteomes, only those which contain ambiguous amino acids were selected for the study, altogether 13 of them (see Appendix A). Based on the types of ambiguous amino acids they have, they were divided into three groups (their Taxonomic ID is stated in the parentheses):

- Group 1, proteomes that contain 'X':
  Salmonella typhi (601), *Arabidopsis thaliana* (3702), *Caenorhabditis elegans* (6239), *Rattus norvegicus* (10116), *Plasmodium falciparum* (36329), *Thermoplasma volcanium* (50339), *Helicobacter pylori* (85962), *Chlamydia pneumoniae* (115711), *Bacillus subtilis* (224308).

- Group 2, proteomes that contain 'B' and 'X':
  Drosophila melanogaster (7227), *Vibrio cholerae* (243277)

- Group 3, proteomes that contain 'B', 'X', and 'Z':
  Homo sapiens (9606), *Mus musculus* (10090)

4.2.2 Methodology

As we know, time efficiency can be significant when dealing with large amounts of data. The proteome files of these 13 model organisms are fairly large ones. Therefore, time efficiency is an important factor in the comparison of the three filtering strategies. To get the total run time (CPU time), $T(A, B)$, which is used when processing data, we need to record and then add times for individual processing steps. In Equation 4.1, T_A is the time to generate A's library; T_B is the time to generate B's library; T_{AB} is the time to use A's library to scan through B's proteome file and then remove the duplicates as well as determine the size of the overlapping file; T_{BA} is the time to use B's library to scan through A's proteome file and then remove the duplicates as well as determine the size of the overlapping file.

$$T(A, B) = T_A + T_B + T_{AB} + T_{BA} \tag{4.1}$$

The total time to complete a task involves many factors such as disk and memory accesses, I/O activities, OS overheads, etc. Therefore, we need to record the time that the processor (CPU) is working only on our programs since multiple processes are running at the same time. $T(A, B)$ is calculated in terms of both "user CPU time" (in user's program) and "system CPU time" (in OS), i.e., the total number of CPU-seconds that the process spent in user mode and system mode respectively. Overhead time for operations such as creating

processes are negligible according to some preliminary experiments. Since the programs are data-intensive and the time for reading and writing data may be significant, system CPU time should be included in the calculation of total time. To lessen the effects of possible timing anomalies, $T(A, B)$ is recorded 10 times and the arithmetic mean is used to calculate a final, reported $T(A, B)$. In general, if one method does a better job of filtering the 9-mers (at the expense of more CPU time) while the increased computational time is less than 50% as compared to the other method, it is considered to be acceptable.

4.2.3 Result

Three tables, one for each filtering method were compiled. The tables give summary information for the 13 model organisms (see Appendix B). They include information about Taxonomic ID, organism name, number of proteins, number of unique 9-mers and number of 9-mer occurrences for each organism.

As described in Section 3.2, two different techniques of counting 9-mers are used. For each technique, three tables are produced according to each filtering strategy. These tables include information about the number of 9-mers common to both organisms, the fraction of unique 9-mers for each organism which are common to both, and the overlap degree between these two organisms. Table 4.2 describes the content of a typical cell (for a pair of proteomes) of one of the two tables. The complete tables of 13×13 organisms are in Appendix C.

Table 4.2: The contents of each portion of a 2×2 unit cell for each pair of proteomes

number of 9-mers which occur in both proteomes	*number of 9-mers which occur in both proteomes ÷ number of 9-mers in proteome of organism in this column*
number of 9-mers which occur in both proteomes ÷ number of 9-mers in proteome of organism in this row	*(number of 9-mers which occur in both proteomes)^2 ÷ (number of 9-mers in proteome of organism in this row × number of 9-mers in proteome of organism in this column)*

One pair of proteomes from each group of proteome files are randomly selected to compare the three strategies. The pairs are: from Group 1, *Helicobacter pylori* (85962) and *Chlamydia pneumoniae* (115711); from Group 2, *Drosophila melanogaster* (7227) and *Vibrio cholerae* (243277); from Group 3, *Homo sapiens* (9606) and *Mus musculus* (10090). The resultant measures including user CPU time, system CPU time and total time are listed in Appendix D.

In Table 4.3, for each pair of proteomes, Method 2 increased the total time by 4.31%, 0.14%, and 0.31% when compared to Method 1. For each pair of proteomes, Method 3 increased the total time by 40.88%, 27.12%, 27.79% separately compared to Method 1. Since Method 2 perturbs the data most, it will not be

used. According to the threshold (i.e., 50%) that we stated in Section 4.2.2, the comparison of timings for Method 1 and Method 3 leads to the conclusion that it takes a reasonable amount of time to do filtering of k-mers containing ambiguities..

Table 4.3: $T(A, B)$ is the total time that is used to process data and count overlaps between two proteomes for all three methods. Both user CPU time and system CPU time are included. Pair 1 is *Helicobacter pylori* (85962) and *Chlamydia pneumoniae* (115711). Pair 2 is *Drosophila melanogaster* (7227) and *Vibrio cholerae* (243277). Pair 3 is *Homo sapiens* (9606) and *Mus musculus* (10090).

Proteomes	**Time $T(A, B)$ (seconds)**		
	Method 1	**Method 2**	**Method 3**
85962, 115711	14.624	15.255	20.603
7227, 243277	292.263	296.412	371.535
9606, 10090	835.921	838.517	1068.244

After comparing the probability of errors among the three filtering methods, we can conclude that method 3 causes the least errors. However, method 3 costs additional computational time to do the filtering because of the manner in which it is implemented. Filtering of k-mers can also be implemented during the overlap determination, rather than as a separate filtration step, with a much lower implementation cost. Implementation via a separate filtration step is used here because the set of k-mers is necessary for statistical analysis. By examining the time efficiency of the three filtering strategies, we can determine whether the filtering is worth the cost and which filtering strategy is more appropriate for each pair of organisms. Table 4.3 shows that the computational cost of the third strategy is within our set threshold compared to method 1. Therefore, the third filtering strategy is used in further stages of the research work.

CHAPTER 5

IMMUNOGENECITY VERSUS SELF(HOST) SIMILARITY

5.1 Proteomic Similarity Analysis

In the last decade, many algorithms have been developed to make use of the linear representation of protein sequence information and search for epitopic motifs [5, 14, 17, 23]. These algorithms search the amino acid sequence of a given protein for characteristics that are believed to be common to antigenic peptides, locating regions that potentially induce cellular immune responses. We are using bioinformatics technology platforms to identify epitopic peptide fragment(s) from disease-associated antigens by following the hypothesis that peptide epitopicity might be regulated by the peptide similarity level to the host's proteome [18, 21, 27]. In order to identify epitopic peptide sequence(s) from disease-associated antigens, it is necessary to consider their similarity to the host's proteome [19, 20, 28]. In Chapter 3, the approximate similarity level between host and pathogen, as well as between different pathogens, is calculated. Overlap between host (usually human) and a pathogen's 9-mer peptides varies from 0.1% - 0.2% of the host's proteome for most bacteria. Overlap between hosts (e.g. human and mouse), as well as between different pathogens, is much higher, ranging from 30% - 40% [3]. Based on this previous work, we are also able to identify regions of low or high self-similarity, host-pathogen similarity, or inter-species similarity (host-host and pathogen-pathogen). In order to determine these regions, expected numbers of matches between two organisms need to be calculated and then compared with the actual numbers of matches between them. If the actual number of matches is lower than expected, such a region is considered to be of *low similarity*; otherwise, the region is of *high similarity*. On the other hand, within the sequences of antigens, there exist regions of high concentration of agretopes, which are recognized by MHC molecules, as well as regions to which antibodies or lymphocytes react. These are known as regions of immunological significance. The theme that is investigated here is whether known regions of immunological significance (e.g. epitopes, agretopes, etc.) come from regions of low host-similarity. If the answer is "yes", it would mean that we are able to use the peptide similarity level to a host's proteome to predict possible regions of immunological significance. The predicted information could therefore be used in vaccine development or development of therapies [3, 18, 21].

Further, pairs of organisms involving human and human viruses are explored. For these pairs, one graph for each k-mer level is generated showing the actual numbers of matches between organisms versus

the expected numbers. At random, one would expect an approximate linear relationship between size of the viral proteome and the number of viral k-mers including repeats occurring in the human proteome. If there are consistently more k-mer occurrences including duplicates in the overlap than what would expect "at random", this means that there is some form of evolutionary pressure applicable to all human viruses which is generating this phenomenon. Such an observation motivates further bench and computational investigation.

5.2 Methodology and Materials

5.2.1 Expected Number of Matches

In order to examine the significance of the level of overlap between two organisms, we need to determine the number of peptides appearing in both organisms under the assumption of random sampling. As a first approximation, we consider the peptides of a given length in a proteome as a random sample (with replacement) from the population of all possible peptides of that length. For k-mers, since each monomer can be any of 20 amino acids, the population size $N = 20^k$ ($N = 20^9 = 5.12{\times}10^{11}$ when $k = 9$). If two samples of size m and n taken randomly with replacement from the set of N objects, the asymptotic behavior is studied when m, n and N all go to infinity, providing that N is much larger than m or n. For each possible object $i = 1, ..., N$, let X_i be the number of times the object is selected in the first sample, and similarly, let Y_i be the number of times the object is selected in the second sample. Then, $E(X_i) = m/N$, $E(Y_i) = n/N$, and since X_i and Y_i are assumed to be independent,

$$E(X_i Y_i) = mn/N^2.$$

As mentioned in Section 3.2, two different techniques of counting the overlaps between two organisms are used. Thus, it is also interesting to study the number of *distinct*, common objects obtained from samples of size m and n. In particular, if A and B are the sets of distinct objects, respectively, then the random variable of interest is the cardinality of intersection between A and B. Call this random variable Z^*. For each possible object $i = 1, ..., N$, let X_i^* be the indicator of the event that object i was selected (at least once) in the first sample

$$X_i^* = 1_{i \in A}$$

and similarly,

$$Y_i^* = 1_{i \in B}.$$

Then we can write

$$Z^* = \sum_{i=1}^{N} X_i^* Y_i^*.$$

Now, object i has m chances of being included in $\boldsymbol{A}$, and the selections are independent. Thus, $E(X_i^*) = \Pr\{i \in A\}$ and similarly, $E(Y_i^*) = \Pr\{i \in B\}$. Under the assumption that X_i^*'s and Y_i^*'s are independent,

we can write

$$E(X_i^* Y_i^*) = \Pr\{i \in A \cap B\} = \Pr\{i \in A\} \times \Pr\{i \in B\}.$$

The expected number of objects in the overlap is then given by

$$E(Z^*) = \sum_{i=1}^{N} E(X_i^*) E(Y_i^*). \tag{5.1}$$

If m is much smaller than N, then $E(X_i^*) \approx E(X_i)$ since the probability that $X_i > 1$ is negligible. More precisely and in general, object i has m independent chances of being excluded from A, each with probability $1 - \frac{1}{N}$, giving a probability of $(1 - \frac{1}{N})^m$ that it does not occur in the selections. Thus,

$$\begin{aligned} E(X_i^*) &= 1 - (1 - \frac{1}{N})^m \\ &= \sum_{j=1}^{m} (-1)^{j-1} \binom{m}{j} N^{-j} \\ &= \frac{m}{N} \left[\sum_{j=1}^{m} (-1)^{j-1} N^{-j+1} \frac{(m-1)!}{j!(m-j)!} \right] \\ &= E(X_i) \left[\sum_{j=0}^{m-1} (-1)^j N^{-j} \prod_{k=1}^{j} \frac{m-k}{1+k} \right]. \end{aligned} \tag{5.2}$$

For large m and N,

$$\begin{aligned} E(X_i^*) &= 1 - (1 - m/(Nm))^m \\ &\approx 1 - \lim_{m \to \inf} (1 - (m/N)/m)^m \\ &= 1 - \exp^{-m/N} \end{aligned}$$

That is, $E(X_i^*)$ can be well approximated by $1 - \exp(-E(X_i))$. If n is also large, then

$$E(Z^*) \approx N(1 - \exp^{-m/N} - \exp^{-n/N} - \exp^{-(m+n)/N}).$$

If m is small but n is large, then $E(Z^*)$ can be approximated by $m(1 - \exp^{-n/N})$.

The error in truncating the alternating series in Equation 5.2 is bounded by the first omitted term. Truncating after the first and second terms thus gives

$$E(X_i)\left(1 - \frac{m-1}{2N}\right) \leq E(X_i^*) \leq E(X_i)\left(1 - \frac{m-1}{2N} + \frac{(m-1)(m-2)}{6N^2}\right).$$

If m is large, we can just as well replace the lower bound by $E(X_i)(1 - E(X_i)/2)$.

If m is appreciably smaller than N, a remarkably good upper bound can be obtained by replacing the product in Equation 5.2 with $((m-1)/2)^j$ (approximating each term of the product by the first term) yielding

$$E(X_i^*) \approx \frac{2m}{2N + m - 1} \approx E(X_i)\left(\frac{1}{1 + E(X_i)/2}\right).$$

The above expression is not necessarily an upper bound for small m, but suffices in practical situations. Thus we have the following boundaries for $E(X_i^*)$:

$$E(X_i)(1 - E(X_i)/2) \leq E(X_i^*) \leq E(X_i)(1 + E(X_i)/2)^{-1} \tag{5.3}$$

The same arguments will apply to $E(Y_i^*) = \Pr\{i \in B\}$.

To find the expected number of unique overlaps, deducing from Equation 5.1 and 5.3 gives

$$\frac{mn}{N}\left(1 - \frac{m}{2N}\right)\left(1 - \frac{n}{2N}\right) \leq E(Z^*) \leq \frac{mn}{N\left(1 + \frac{m}{2N}\right)\left(1 + \frac{n}{2N}\right)}$$

providing that $m \ll N$ and $n \ll N$. By ignoring terms of $O(mn/N^2)$, we get the simpler bounds

$$\frac{mn}{N}\left(1 - \frac{m+n}{2N}\right) \leq E(Z^*) \leq \frac{mn}{N + \frac{m+n}{2}}.$$

Suppose that the total number of occurrences in the second sample (i.e.,"HOST") of objects occurring at least once in the first sample (i.e.,"BUG") needs to be determined. Call this random variable Z. Then, we get

$$E(Z) = \sum_i E(X_i^*)E(Y_i) \approx \frac{mn}{N + m/2} \tag{5.4}$$

by using the right-hand approximation in Equation 5.3. The number of expected matches between two organisms is calculated using Equation 5.4 during further stages of the research work. The above method was contributed by Dr. Mik Bickis in the Department of Mathematics and Statistics [1].

5.2.2 Epitopes

Compared to bacteria, well-studied human viruses contain more complete information of known epitopes and thus are better candidate proteomes. Therefore, *Human immunodeficiency virus 1* (HIV-1) (Taxonomic ID: 11676) is selected to conduct the similarity analysis. The list of known epitopes which is used in the analysis contains only $CD8^+$ epitopes. That is, these epitopes are only known to be recognized by MHC-I molecules. For HIV-1, the summary of known epitopes is downloaded from a database at Los Alamos National Laboratory (LANL) (http://www.hiv.lanl.gov/content/immunology/tables/tables.html). The list of all HIV $CD8^+$ epitopes mapped to within a region of 21 amino acids or less is used to identify the regions of immunological significance.

5.2.3 Host-similarity

As mentioned in Chapter 3, since most viruses have few overlaps with hosts at the 8-mer or 9-mer level (too few to allow for statistically valid results), overlap at the 7-mer level is explored for the first theme. Using lengths of 7 amino acids is acceptable since the T-cell immune system is known to also recognize amino sequences of this length [7, 9]. The overlaps between HIV-1 and human proteome are counted and

compared with the expected number of matches, then the regions of low host-similarity are determined. For each position in the viral proteome, if the actual number of matches is less than the expected number of matches or the actual number of matches is zero, it should be included in the regions of low host-similarity.

Once both regions are identified, a graph is generated to compare the regions of low host-similarity and regions of immunological significance (i.e., epitopes). The overlap information as well as the number of epitopes for each position within the viral proteome is plotted. By graphing the above two sets of information, a comparison between the regions of low host-similarity and the regions of known epitopes is produced.

Correlation of the two data sets $y1$ and $y2$ in the graph, which is used to essentially tell how close the linearly related two data sets $y1$ and $y2$ are, is measured by "correlation coefficient" (r). The coefficient r is a measure of the strength of the relationship. If two data sets $y1$ and $y2$ fell exactly on a line with negative slope, $r = -1$. In this case we say that $y1$ and $y2$ are "perfectly negatively linear correlated". If $r = 0$, we say that $y1$ and $y2$ are "uncorrelated". In general, $r < -0.5$ indicates $y1$ and $y2$ are approximately negatively linear correlated. Also, a p-value is used for testing the hypothesis of no correlation. It is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero. If p-value is small, say less than 0.05, then the correlation r is significant.

5.2.4 Actual Overlaps versus Expected Overlaps

For pairs involving human and human viruses, one graph at each k-mer level (where k = 5,6,7,8,9) is generated showing the actual numbers of matches between organisms versus the expected numbers. These results have been determined for 24 human viruses (listed in Appendix E). The length of these viral proteomes varies from 1000 to 7000 amino acids. Expected number of matches including duplicates between each pair of proteomes is calculated using Equation 5.4. Since the size of the viral proteome $m \ll N$ in each case, one would expect an approximate linear relationship between size of the viral proteome and the number of viral k-mers including repeats occurring in the human proteome. That is, we get $E(Z) \approx \frac{mn}{N}$ from Equation 5.4 providing $m \ll N$. Also, n and N is constant at each k-mer level. Therefore, $E(Z)$ is expected to be approximately linearly related to m. Again, correlation of the two data sets x and y along each line in the graph is measured by correlation coefficient (r). If two data sets x and y fell exactly on a straight line with positive slope, then $r = 1$; while if they fell exactly on a line with negative slope, $r = -1$. In these cases we say that x and y are "perfectly positively linear correlated" or "perfectly negatively linear correlated". If $r = 0$, we say that x and y are "uncorrelated". In general, $r > 0.5$ indicates x and y are approximately linear correlated. Also, a p-value less than 0.05 indicates that the correlation r is significant.

5.3 Results and Discussion

5.3.1 Proteomic Similarity Analysis

The HIV-1 proteome consists of 3517 occurrences of 3079 unique 7-mers. Using Equation 5.4, the expected number of overlaps between HIV-1 and human is

$$E(Z_{HIV-1}) \approx \frac{mn}{N + m/2} = \frac{3517 \times 16171995}{1.28 \times 10^9 + \frac{3517}{2}} \approx 44.$$

Then, for each 7-mer in HIV-1 proteome, the expected number of overlaps is $44 \div 3079 = 0.0143$. Therefore, if the actual number of overlaps is zero (< 0.0143) for any position in the HIV-1 proteome, it is considered to be included in the regions of low host-similarity. The upper graph in Figure 5.1 indicates the overlapping information between the HIV-1 and human proteomes for each position in the HIV-1 proteome. The lower graph in Figure 5.1 indicates the known epitopes of HIV-1 for each position in the HIV-1 proteome. By calculating the correlation coefficient r and the p-value, we have $r = -0.0029$ (> -0.5) and p-value is 0.8819 (> 0.05). Therefore, results show that the correlation between regions of low host-similarity and regions of known epitopes for HIV-1 proteome at the 7-mer level is not statistically significant. One problem is that there are too many non-overlaps along the sequence of the viral proteome. Thus it is hard to distinguish between the regions of low host-similarity and the regions of high host-similarity from Figure 5.1. In an attempt to circumvent this problem, the comparison within only one viral protein in HIV-1 proteome is performed. The Gag polyprotein is selected since it contains a relatively large number of epitopes. In Figure 5.2, both number of epitopes and number of overlaps for each position in the protein Gag polyprotein are indicated. Again, we have $r = -0.0426$ (> -0.5) and p-value is 0.3395 (> 0.05). Thus, results show that the correlation between these two regions is not statistically significant either.

5.3.2 Actual Overlaps Versus Expected Overlaps

One graph at each k-mer level (where $k \in [5..9]$) is generated showing the actual numbers of matches between organisms versus the expected numbers. In these graphs, the x-axis represents the sizes of the peptide universes of the 24 viral proteomes, while the y-axis represents the count of viral k-mer occurrences including duplicates in the human proteome. For each viral proteome, overlapping occurrences including repeats, number of involved human proteins, unique overlaps and expected overlapping occurrences including repeats are plotted. According to the model in 5.3.1, one would expect an approximately linear relationship between the size of the viral proteome and the number of viral k-mers including repeats occurring in the human proteome. All graphs show that there are consistently more k-mers in the overlap than what we would expect "at random". The correlation coefficients r as well as the p-values for each linear regression lines in each graph are listed in Table 5.1. At the 5-mer level, for data sets of all four measurements in Figure 5.3, the correlation coefficient $r \in (0.9..1)$ and the p-value is far less than 0.05. This means that the linear

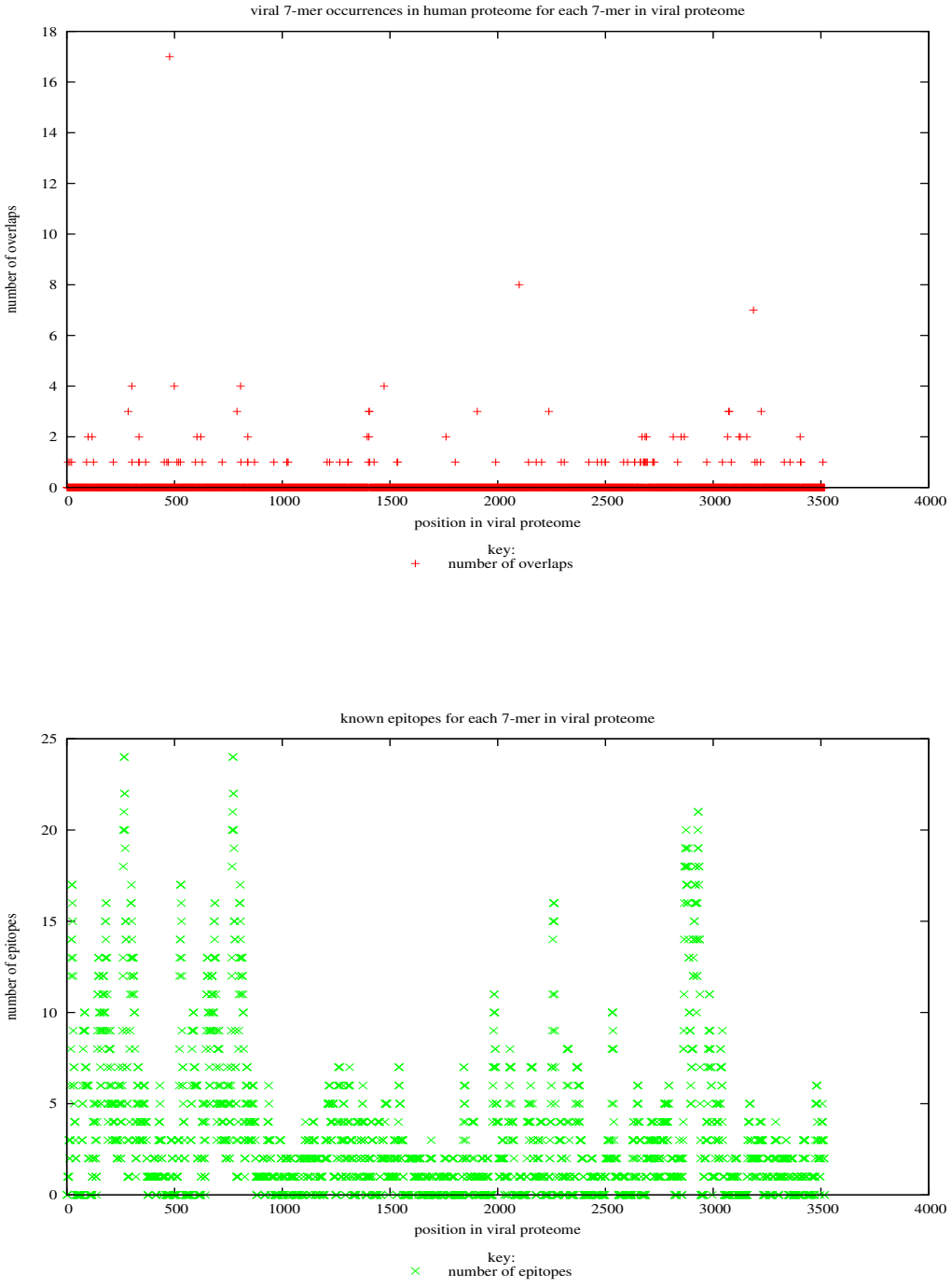

Figure 5.1: Two graphs showing counts of matches and counts of epitopes for each 7-mer in the viral proteome. The x-axis represents the position of each 7-mer in HIV-1 in both graphs. The y-axis represents the count of viral 7-mer occurrences in the human proteome in the upper graph and the count of known epitopes in the lower one.

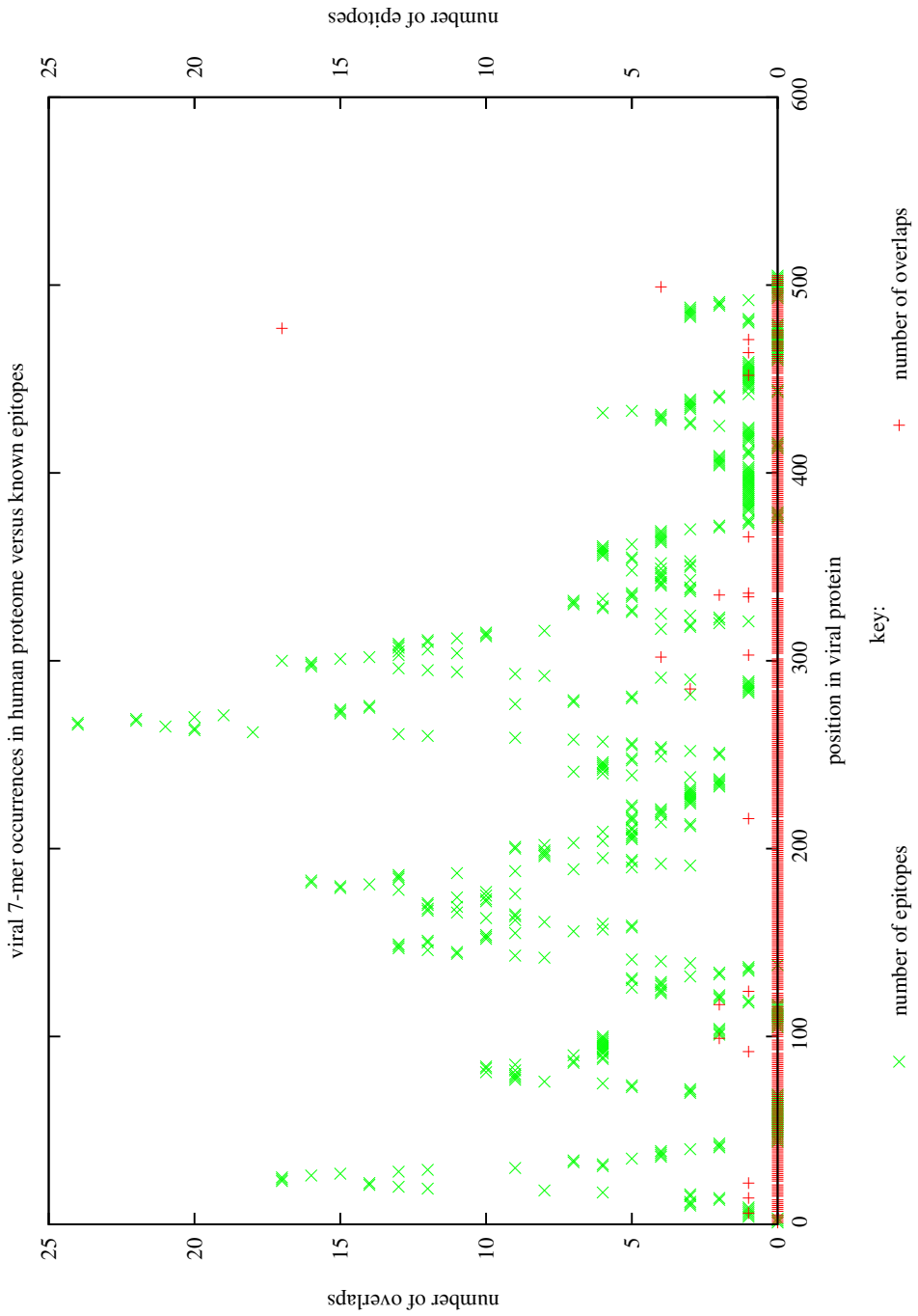

Figure 5.2: A graph showing counts of matches versus counts of epitopes for each 7-mer in the HIV-1 Gag polyprotein. The x-axis represents the position of each 7-mer in HIV-1 Gag polyprotein, while the left-hand y-axis represents the count of viral 7-mer occurrences in the human proteome and the right-hand y-axis represents the count of known epitopes.

correlation of data sets in each measurement is extremely significant. At the 6-mer level, the correlation coefficient $r \in (0.8..1)$ and the p-value is less than 0.05 for data sets of all four measurements in Figure 5.4. Thus, the linear correlation of data sets in each measurement is highly significant. At the 7-mer level, the correlation coefficient $r \in (0.6..1)$ and the p-value is less than 0.05 for data sets of all four measurements in Figure 5.5. Again, the linear correlation of data sets in each measurement is significant. Also, the expected overlapping occurrences including repeats at each k-mer level is linearly correlated. However, at the 8-mer and 9-mer levels, the correlation coefficients r of the actual overlapping occurrences including repeats are all less than 0.5 and the p-values are larger than 0.05. This means that the linear correlations for them are not significant. Similarly, the linear correlation of data sets in neither the number of involved human proteins nor the unique overlaps at the 8-mer and 9-mer levels is significant.

In each graph, noteworthy viral proteomes are labeled with their accession number: *Hepatitis C virus* (P26663), *Infectious hematopoietic necrosis virus* (X89213), *Crimean-Congo hemorrhagic fever virus* (Q52NX4), *Lake Victoria marburgvirus* (Z12132), *Human coronavirus strain SARS* (P59641). These viral proteomes tend to have higher values than the linear regression lines at each k-mer level.

Table 5.1: Summary of the correlation coefficients r and p-values for each linear regression lines in Figure 5.3, Figure 5.4, Figure 5.5, Figure 5.6 and Figure 5.7.

k-**mers**	**actual overlapping occurrences including repeats**	
	r	p-**value**
5-mer	0.9288	5.9044E-11
6-mer	0.8002	2.6741E-6
7-mer	0.5719	0.0035
8-mer	0.0906	0.6736
9-mer	-0.1138	0.5966

k-**mers**	**# of human proteins involved in overlap**	
	r	p-**value**
5-mer	0.9087	8.2576E-10
6-mer	0.8603	7.0204E-8
7-mer	0.6466	0.0006
8-mer	0.1560	0.4665
9-mer	0.0832	0.6991

k-**mers**	**actual unique overlaps**	
	r	p-**value**
5-mer	0.9995	1.9302E-34
6-mer	0.9780	1.8366E-16
7-mer	0.8725	2.7423E-8
8-mer	0.2288	0.2822
9-mer	-0.1420	0.5081

k-**mers**	**expected overlapping occurrences including repeats**	
	r	p-**value**
5-mer	0.9999	7.4658E-61
6-mer	0.9999	4.6644E-62
7-mer	0.9999	4.1053E-41
8-mer	0.9999	6.4987E-70
9-mer	0.9999	4.1854E-49

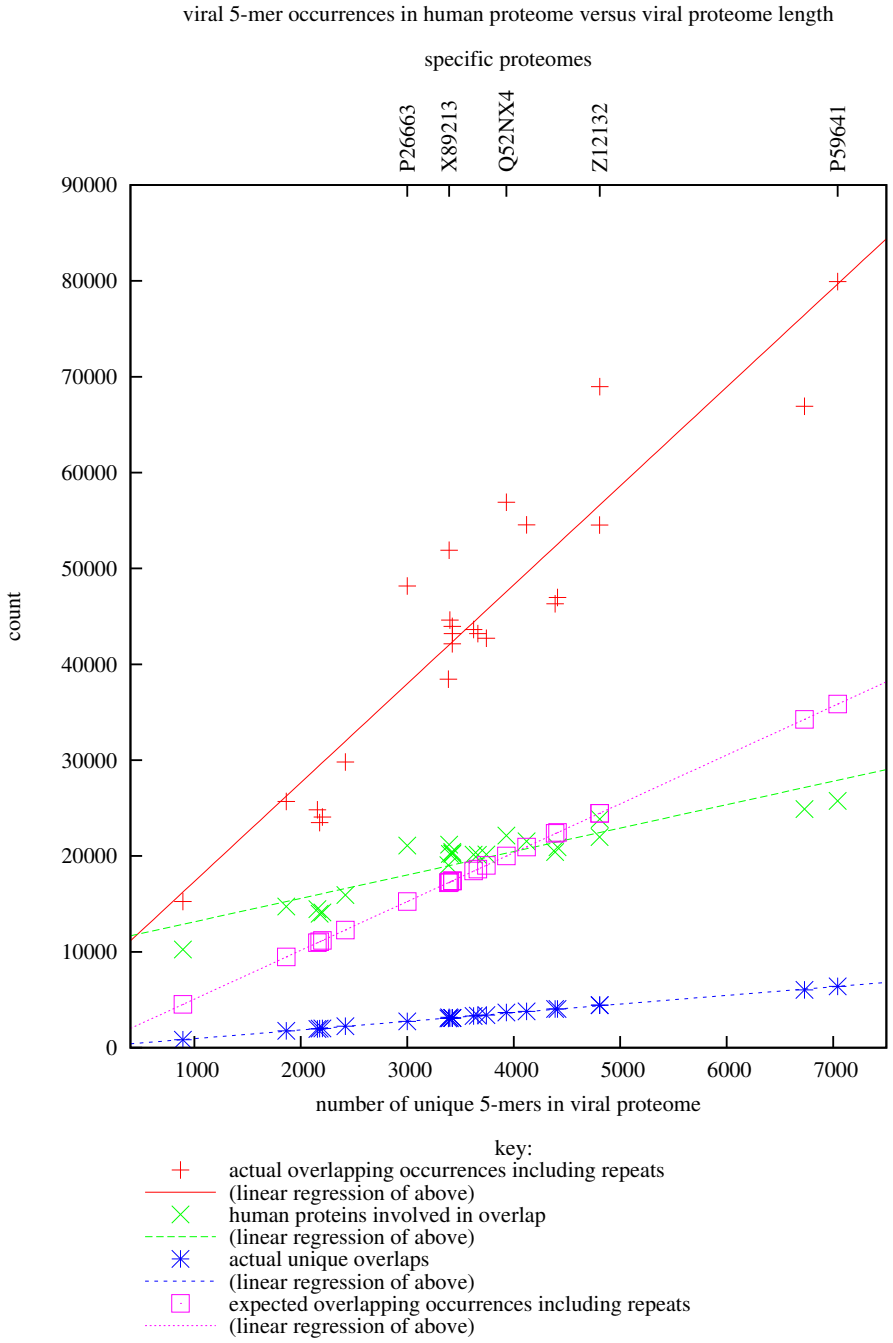

Figure 5.3: A graph showing counts of matches between 24 viruses and human at the 5-mer level. Shown are overlapping occurrences, unique overlaps, expected overlapping occurrences and number of human proteins involved. The x-axis represents the length of the various viral proteomes, while the y-axis represents the count of viral 5-mer occurrences in the human proteome or the count of involved human proteins.

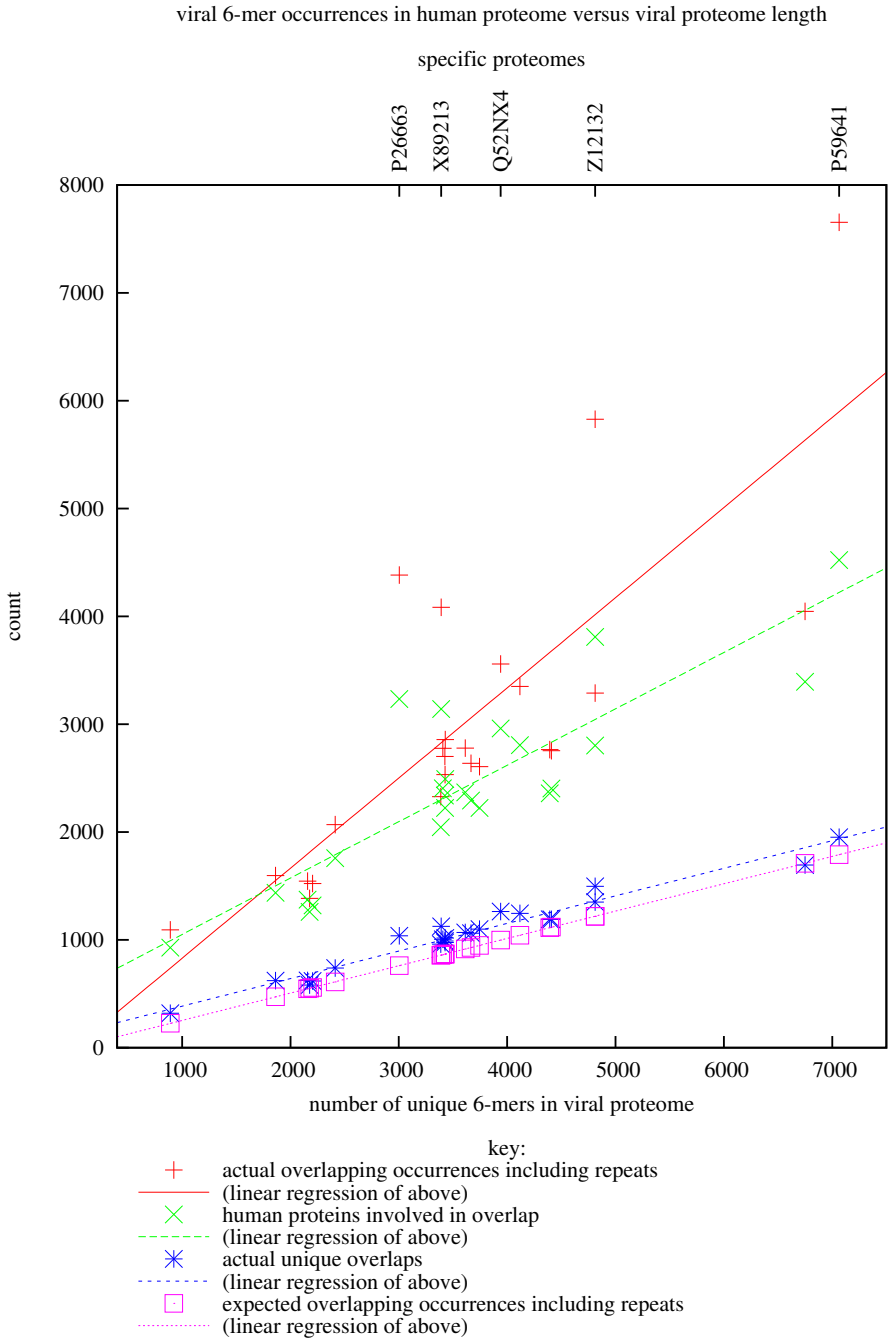

Figure 5.4: A graph showing counts of matches between 24 viruses and human at the 6-mer level. Shown are overlapping occurrences, unique overlaps, expected overlapping occurrences and number of human proteins involved. The x-axis represents the length of the various viral proteomes, while the y-axis represents the count of viral 6-mer occurrences in the human proteome or the count of involved human proteins.

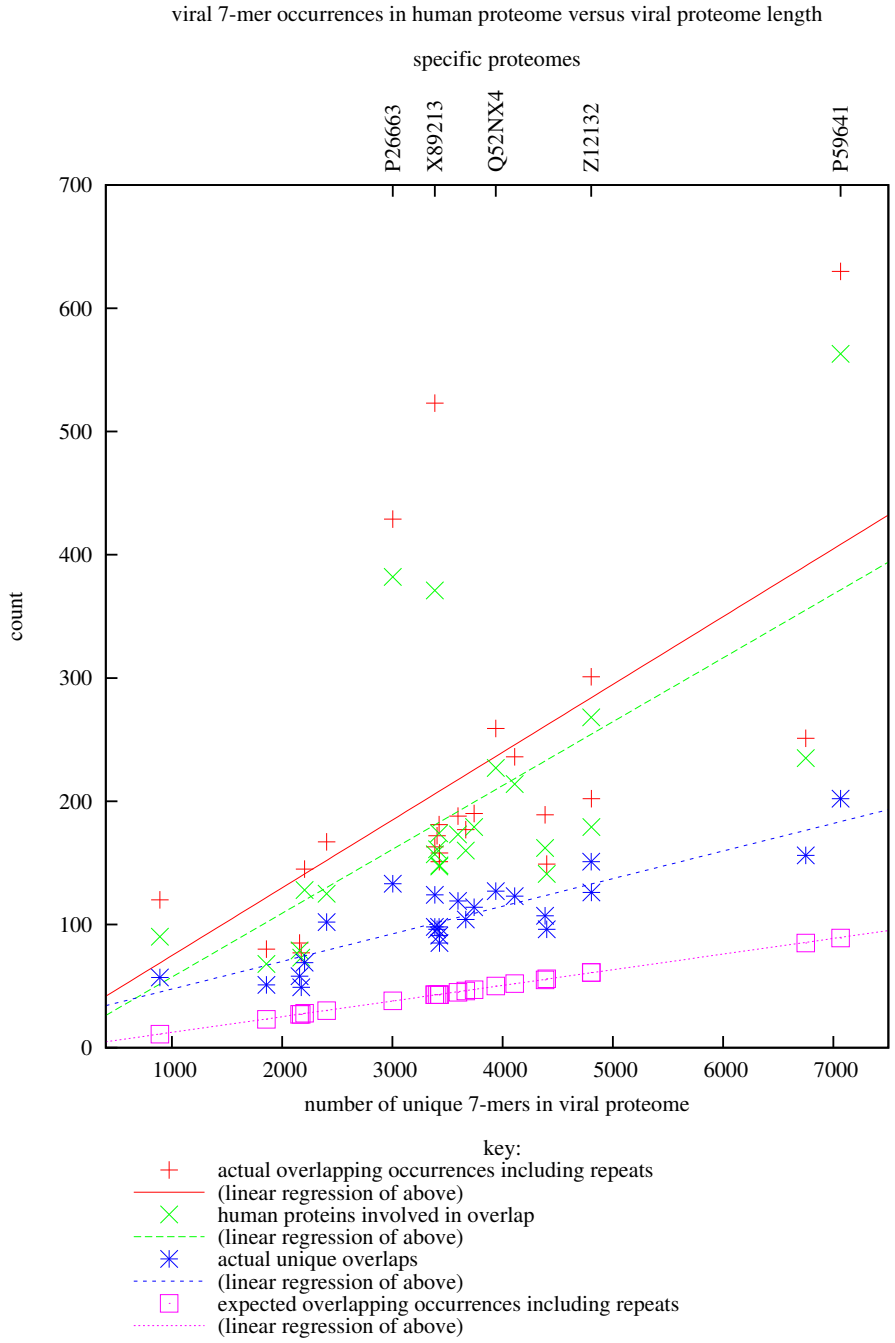

Figure 5.5: A graph showing counts of matches between 24 viruses and human at the 7-mer level. Shown are overlapping occurrences, unique overlaps, expected overlapping occurrences and number of human proteins involved. The x-axis represents the length of the various viral proteomes, while the y-axis represents the count of viral 7-mer occurrences in the human proteome or the count of involved human proteins.

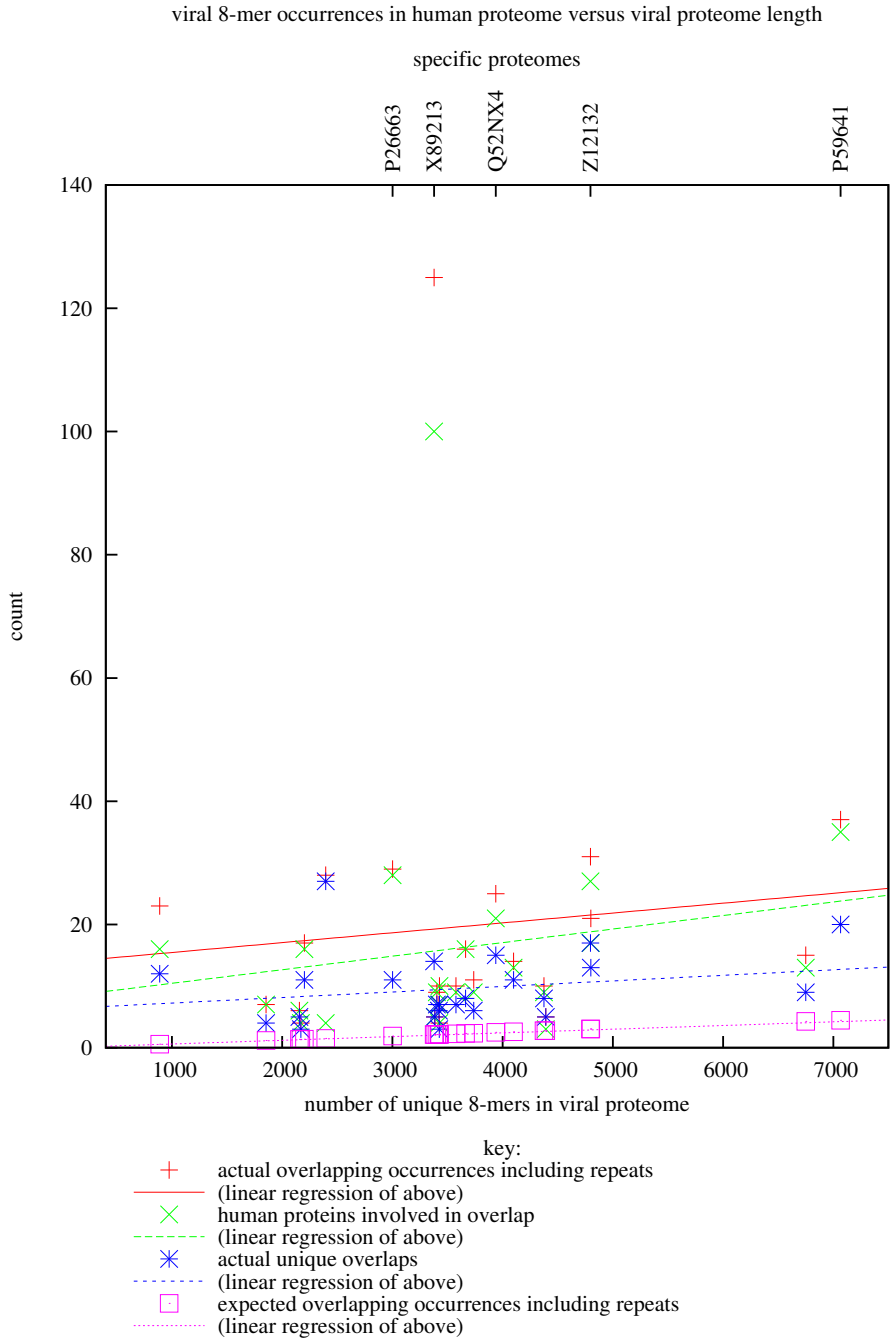

Figure 5.6: A graph showing counts of matches between 24 viruses and human at the 8-mer level. Shown are overlapping occurrences, unique overlaps, expected overlapping occurrences and number of human proteins involved. The x-axis represents the length of the various viral proteomes, while the y-axis represents the count of viral 8-mer occurrences in the human proteome or the count of involved human proteins.

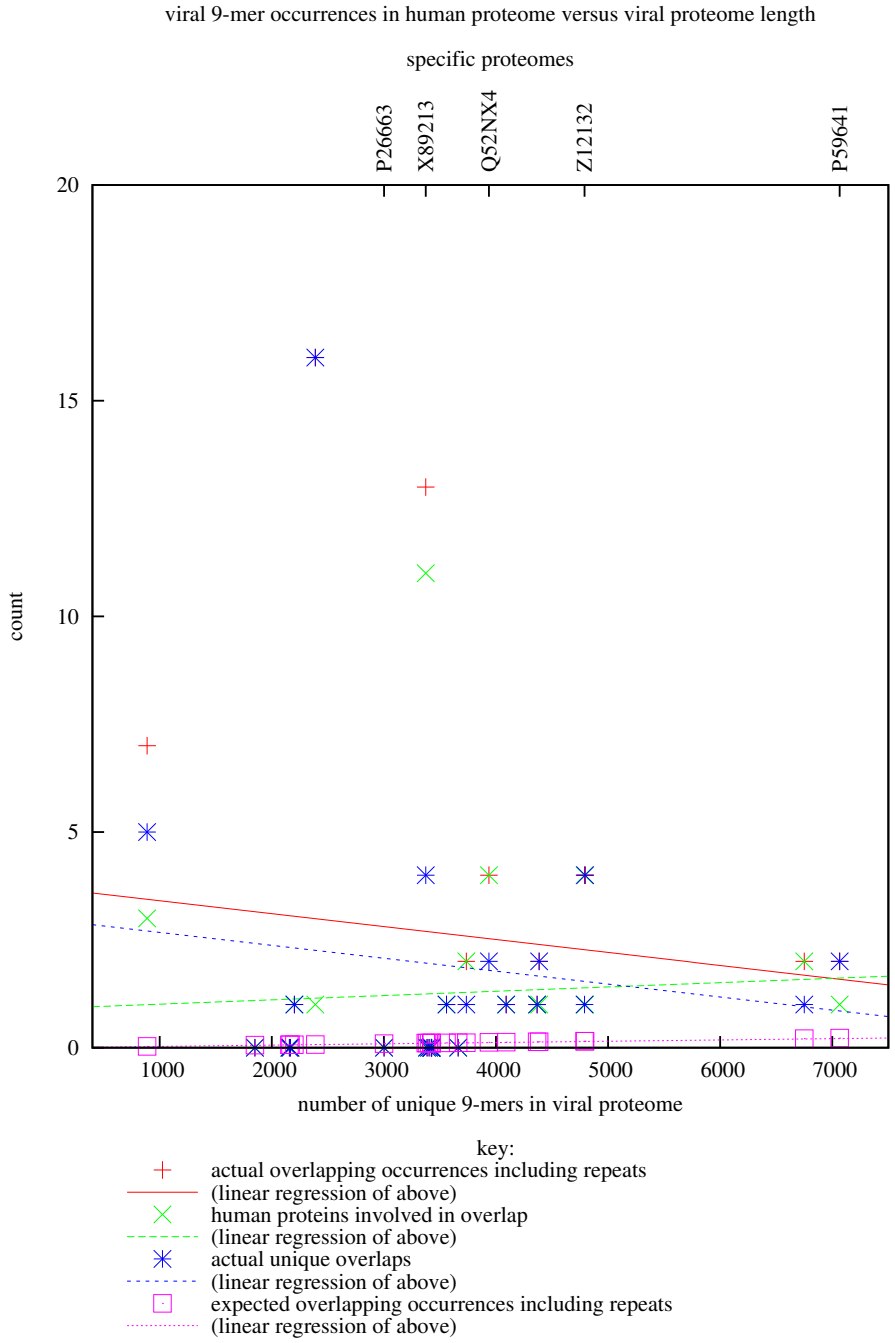

Figure 5.7: A graph showing counts of matches between 24 viruses and human at the 9-mer level. Shown are overlapping occurrences, unique overlaps, expected overlapping occurrences and number of human proteins involved. The x-axis represents the length of the various viral proteomes, while the y-axis represents the count of viral 9-mer occurrences in the human proteome or the count of involved human proteins.

CHAPTER 6

LOCALITY CLUSTERING

6.1 Structure In Overlap/Non-overlap Location

A detailed investigation of the overlaps between host and pathogenic species may reveal that the distribution of the locations of these overlaps within either the pathogenic proteome or host proteome has structure. Therefore, another question that this thesis pursues is whether the locality clustering is statistically significant. Here "structure" could refer to various aspects of overlaps such as locality clustering, tissue clustering, etc. These regions are distinguished by having a level of overlap which varies significantly from what one would expect assuming k-mers are distributed at random throughout the proteomes. Regions with significantly more or fewer overlaps (i.e. both high and low similarity) are noteworthy. Host-similarity refers to the overlap between pathogen and host (e.g. human or mouse). One could investigate structure in occurrences of low host-similarity, high host-similarity, or both. This work investigates structures in both cases. Self-similarity refers to repeated patterns in the proteome of a single organism. It is relevant in the study of autoimmune disease. A chi-square analysis is employed to analyze locality clustering. The chi-square analysis techniques are described in detail in Section 6.2.2.

6.2 Chi-square Analysis of Structure

6.2.1 Summary of Viral Proteomes and Human Proteome

Host-similarity

Table 6.1 is the summary information of three viral proteomes (*HIV-1* (Taxonomic ID:11676), *HIV-2* (Taxonomic ID:11709) and *Influenza A virus* (Taxonomic ID:93838)) and *Homo Sapiens* (Taxonomic ID:9606).

Table 6.2 is the summary of overlaps between each viral proteome (HIV-1, HIV-2 and Influenza A virus) and human proteome.

Table 6.1: Summary information of three viral proteomes (HIV-1, HIV-2 and Influenza A virus) and human proteome at the 5-, 6- and 7-mer level.

Taxonomic ID	**Name**	**5-mer**		
		# of proteins	**# of unique 5-mers**	**# of 5-mer occurrences**
9606	Homo sapiens	37991	2388563	16249364
11676	HIV-1	9	3082	3535
11709	HIV-2	9	3285	3723
93838	Influenza A virus	10	4412	4427
Taxonomic ID	**Name**	**6-mer**		
		# of proteins	**# of unique 6-mers**	**# of 6-mer occurrences**
9606	Homo sapiens	37991	8247275	16210640
11676	HIV-1	9	3084	3526
11709	HIV-2	9	3283	3714
93838	Influenza A virus	10	4408	4417
Taxonomic ID	**Name**	**7-mer**		
		# of proteins	**# of unique 7-mers**	**# of 7-mer occurrences**
9606	Homo sapiens	37991	10431975	16171995
11676	HIV-1	9	3079	3517
11709	HIV-2	9	3278	3705
93838	Influenza A virus	10	4400	4407

Table 6.2: Summary of overlaps between each viral proteome (HIV-1, HIV-2 and Influenza A virus) and human proteome at the 5-, 6- and 7-mer level.

Viral Proteome	**5-mer**		
	# of unique overlaps	**# of overlapping occurrences**	**# of human proteins involved**
HIV-1	2792	36640	18158
HIV-2	2949	47099	19617
Influenza A virus	4036	46966	20909
Viral Proteome	**6-mer**		
	# of unique overlaps	**# of overlapping occurrences**	**# of human proteins involved**
HIV-1	904	2181	1873
HIV-2	1006	4746	2862
Influenza A virus	1191	2754	2404
Viral Proteome	**7-mer**		
	# of unique overlaps	**# of overlapping occurrences**	**# of human proteins involved**
HIV-1	96	167	152
HIV-2	121	1350	536
Influenza A virus	96	149	141

Self-similarity

Table 6.3 is the summary information of human protein desmoglein-3 (Dsg3) and human proteome without the desmoglein-3 human protein.

Table 6.3: Summary information of human protein desmoglein-3 (Dsg3) and human proteome without the desmoglein-3 human protein at the 5-, 6- and 7-mer level.

Taxonomic ID	**Name**	**5-mer**		
		# of proteins	**# of unique 5-mers**	**# of 5-mer occurrences**
9606 (except Dsg3)	Homo sapiens	37990	2388524	16248369
Dsg3	Desmoglein-3	1	992	995
Taxonomic ID	**Name**	**6-mer**		
		# of proteins	**# of unique 6-mers**	**# of 6-mer occurrences**
9606 (except Dsg3)	Homo sapiens	37990	8246693	16209646
Dsg3	Desmoglein-3	1	994	994
Taxonomic ID	**Name**	**7-mer**		
		# of proteins	**# of unique 7-mers**	**# of 7-mer occurrences**
9606 (except Dsg3)	Homo sapiens	37990	10431100	16171002
Dsg3	Desmoglein-3	1	993	993

Table 6.4 is the summary of overlaps between the human protein desmoglein-3 (Dsg3) and human proteome without the desmoglein-3 human protein.

Table 6.4: Summary of overlaps between the human protein desmoglein-3 (Dsg3) and human proteome without the desmoglein-3 human protein at the 5-, 6- and 7-mer level.

Viral Proteome	**Dsg3**		
	# of unique overlaps	**# of overlapping occurrences**	**# of human proteins involved**
5-mer	953	17700	10134
6-mer	412	2550	1494
7-mer	118	468	270

6.2.2 Methodology

Since regions with significantly more or fewer overlaps (i.e. both high and low similarity) are noteworthy in the chi-square analysis, we need to calculate the expected number of matches for each region and then

compare it with the corresponding observed number. If the observed number of matches is lower than the expected number for the region, it is considered as low similarity. Otherwise, higher than expected indicates high similarity. Given a single, linear sequence, the peptide universes and overlaps between the pair of proteomes are generated. Then the single long proteomic sequence for each organism is divided into evenly-sized segments. Each segment should have an expected number of non-overlaps/overlaps of 5 or more. Therefore, no more than $E/5$ (where E is the expected number of non-overlaps/overlaps) segments will be made and they will be merged into bigger segments if necessary. In the case of 5-mers, non-overlapping 5-mers are considered since there exists more overlaps than non-overlaps at the 5-mer level. However, with 6-mers and 7-mers, there are more non-overlapping k-mers than overlapping ones. Therefore, it makes more sense to look at the distribution of the locations of 6-mers and 7-mers which also occur in human (i.e., overlaps). Thus in the locality clustering analysis, when we talk about "overlap", we actually mean non-overlap in the 5-mer case and overlap in the 6-mer and 7-mer cases.

Suppose a "reasonable" number of overlaps (i.e., at least 10) exist between a pair of organisms. Then a chi-square test can be used to analyze whether the structure (clustering) in the regions of low or high host-similarity or self-similarity is statistically significant. Two types of chi-square analysis are used in the work. One technique is to do the chi-square analysis by dividing the viral proteome into evenly-sized segments. The other is to do the chi-square analysis by dividing the multi-protein viral proteome into segments of individual proteins. That is, the segment boundary is the same as the protein boundary.

For the first technique, viral proteome is divided into evenly-sized segments. The expected number of overlaps per segment is E_i = ((# of overlaps)$\div$(# of k-mers))$\times$(segment size) for each segment i except for the last segment may have a longer length and a different expected overlaps. In the 5-mer and 6-mer cases, viral proteomes are broken up into segments of 100 amino acids. In the 7-mer case, if we break up the viral proteome in the same way as 5-mer and 6-mer, the expected number of overlaps per segment might be less than 5 because the number of overlap occurrences and overlapping unique k-mers are significantly lower for 7-mers than for the 5-mers and 6-mers. The "rule of thumb" is that the expected number per segment should be at least 5. Therefore, for 7-mers we break up the viral proteome into segments with length of longer than 100 amino acids: 200 for HIV-1, 150 for HIV-2, 250 for Influenza A virus. In general, we want the p-value to be small in order to find significance. Having a small p-value could be the following two cases: the data set is significant, or the χ^2 statistic is not approximate. Therefore, there is a trade-off involved: for statistical significance, one wants lots of segments with small length to make the p-value small in order to find significance; for precision, one wants fewer, large segments which have a larger expected number of overlaps to make the χ^2 statistic to be approximate. Thus, for HIV-1, we have $3517 \div 200 \approx 18$ segments and $E_i = (112 \div 3517) \times 200 \approx 6.37$ per segment; for HIV-2, we have $3705 \div 150 \approx 25$ segments and $E_i = (138 \div 3705) \times 150 \approx 5.587$ per segment; for Influenza A virus, we have $4407 \div 250 \approx 18$ segments and $E_i = (96 \div 4407) \times 250 \approx 5.446$ per segment. The null hypothesis is: "the locations of the observed overlapping k-mers (where k=5,6,7) are distributed at random amongst the segments; i.e., the

locality clustering to the distribution of the overlapping k-mers is not significant." Next,

$$\chi^2 = \sum_{segment\ i} \frac{(O_i - E_i)^2}{E_i} \tag{6.1}$$

will be calculated for each segment where E_i is the expected number of matches and O_i is the observed number of matches. The degrees of freedom for the calculated χ^2 value is one less than the number of segments. In order to use the χ^2 value to determine whether or not there is clustering in certain regions, the calculated chi-square value is compared with the values in a standard chi-square table. If the calculated value is greater, one can reject the null hypothesis and conclude that the number of overlaps appearing in the segments is different from random. Further, following conventional criteria, p-values less than 0.05 will be taken as statistically significant. For a viral proteome, if the locations of overlaps are distributed everywhere along the proteomic sequence, no clustering is found, or alternatively, we have a large cluster which includes the whole proteome.

In the above method, we can consider segments right up to the length (m) of the viral proteome (i.e., the last k-mer starts at position $m - k + 1$). That is, the expected number of overlaps per segment for each viral proteome is ((# of overlaps)$\div$(# of k-mers in viral proteome))$\times$(length of each segment), except for the last segment which has an expected value of $E_i\times$(proportion of last segment). Then, the chi-square analysis is performed except that for the last segment in the viral proteome, different value of E_i is used.

The above χ^2 analysis is simpler if the proteome of each organism is a single, linear sequence. This is already the case for viruses which have a polyprotein. However, for other organisms, their (multiple) protein sequences result in a procedural problem. Some segments in the viral proteome may include one portion from one viral protein and the other portion from another viral protein. If such a segment is determined to be statistically significant, it is hard to tell which portion of the segment have clustering, or all portions of the segment have clustering. To eliminate the problem, protein boundaries could be considered as segment boundaries when dividing the viral proteome into segments.

A second chi-square analysis technique, in which each individual protein of an organism is analyzed, is proposed to circumvent the "artificial peptide" problem described above. Instead of dividing the linearized proteome into equally-spaced segments, one can make the segment boundaries correspond to protein boundaries. Some of the proteins in the proteomes of the multi-protein organisms are small and have fewer than 5 expected overlaps. These proteins will not be included in the second type of analysis. Then the first type of analysis is repeated by using the Equation 6.1. The degrees of freedom are still one less than the number of segments. The calculated χ^2 values are then compared to tabulated values and a conclusion drawn on the existence of clustering in the locations of low or high host-similarity. Again a p-value of less than 0.05 will be taken as indicating a statistically significant result. For organisms which have a polyprotein (i.e., human protein Desmoglein-3), the second technique is unnecessary since there will be only one segment to be analyzed. For organisms with a multiple-protein proteome (i.e., *HIV-1*, *HIV-2*, *Influenza A virus*), the second type of analysis is performed and the results from the two techniques are compared. There is a short-coming

of the second technique: for a given organism, some areas of its proteome may be subjected to only the first type of analysis.

It may occur that structure or clustering occurs preferentially with certain lengths of peptides. For the vertebrate immune system, the typical length of peptides presented to $CD8^+$ T cells by MHC class I molecules is nine residues. Peptides of 7 or 8 amino acids are also presented, but not with the same preference. However, after investigating the generated overlaps between selected viral proteomes and human, we observed that there are too few overlaps at the 8-mer and 9-mer levels to conduct the χ^2 analysis (see Table F.4 and F.5 in Appendix F). Therefore, the clustering analysis most relevant to the T-cell immune system is that performed at the 7-mer level. The B cell immune system can recognize peptides of 5 or 6 residues. Thus, Observations of clustering at the 5-mer or 6-mer level may help to identify regions of immunological significance in the B-cell immune system.

If it is found that the overlaps are not distributed at random using the χ^2 analysis, the following question naturally arises: are there any particular segments which have a statistically significant greater or lower number of overlapping k-mers than expected? This question can be answered by analyzing each row of tables in Appendix G. If $(O_i - E_i)^2/E_i \geq 4$ for row i, then we can conclude that the particular observation is different from random with statistical significance. In the case of 5-mers, analysis considers extent of non-overlaps and having significantly fewer non-overlaps than expected corresponds to being "humanlike", while having more non-overlaps than expected corresponds to being "non-humanlike". However, in the cases of 6-mers and 7-mers, analysis consider extent of overlaps and having significantly more overlaps than expected corresponds to being "humanlike", while having fewer overlaps than expected corresponds to being "non-humanlike". For well-studied areas of the viral proteomes, agreement between these regions being humanlike/non-humanlike with their known characteristics needs to be validated. For areas of the viral proteomes which are less studied or for which fewer biochemical results are available, the analyzed results of these regions being humanlike/non-humanlike can be used to predict their biological functions or attribute. The predictions may be useful in developing therapies for the diseases caused by the viruses or to help life scientists understand the mechanisms of the diseases.

Similar chi-square analysis can be performed regarding the self-similarity of human protein Desmoglein 3 (Dsg3) to the remaining human proteome. Pemphigus is an autoimmune disorder that causes blistering and raw sores on skin and mucous membranes. As with other autoimmune disorders, it is caused when the body's defenses mistake its own tissues as foreign, and attack the cells. The most common form of the disorder is pemphigus vulgaris. It occurs when antibodies attack Desmoglein 3, a protein that keeps cells bound together. Thus, cells simply fall apart, causing skin to slough off. By investigating the overlaps between protein Dsg3 and other proteins in the human proteome, we can determine whether or not there is structure (clustering) to the regions of self-similarity. Then certain regions of Dsg3 could be determined whether or not being humanlike/non-humanlike. Our previous methodology can be used, with the human protein Desmoglein 3 considered as "BUG" proteome while the human proteome except protein Desmoglein 3 is

considered as the "HOST" proteome. Since protein Dsg3 contains only 999 amino acids and has no sub-protein, we only perform the chi-square analysis by dividing Dsg3 into regions of certain length. All of 5-mer, 6-mer and 7-mer level analysis are performed and comparisons of the χ^2 value with the values in a standard chi-square table are conducted. Again a p-value of less than 0.05 will be taken as indicating a statistically significant result.

6.2.3 Clustering Analysis and Results

The executive summary is that there is statistically significant clustering in the location of the overlapping/non-overlapping k-mers. Note that in the second type of clustering analysis, the clustering (or non-clustering) involves only certain proteins of a multi-protein organism.

Chi-square Analysis

A chi-square analysis on the distribution of overlapping/non-overlapping k-mers within each viral proteome (HIV-1, HIV-2 and Influenza A Virus) are used by employing two different techniques. In the case of 5-mers, non-overlapping 5-mers are considered since there exists more overlaps than non-overlaps at the 5-mer level. However, with 6-mers and 7-mers, there are more non-overlapping k-mers than overlapping ones. Therefore, it makes more sense to look at the distribution of the locations of 6-mers and 7-mers which also occur in human (i.e., overlaps). The detailed summary information about the distribution of these non-overlaps/overlaps is listed in Appendix G.

Table 6.5 contains the summary information of the calculated χ^2 value, degrees of freedom (d.f.) and p-value for each analysis. By conventional criteria, we can conclude that the result of each analysis except the 7-mer of HIV-1 with technique 2 is considered statistically significant. That is, the number of overlaps appearing in the segments is different from random. We reject the corresponding null hypothesis and conclude that there is structure (clustering) to the location of overlapping k-mers within the viral proteomes.

Similarity to Human Proteome

Table 6.6, Table 6.7 and Table 6.8 contain information from analyzing the similarity to the human proteome for HIV-1 at the 5-mer, 6-mer and 7-mer level respectively. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. Summarizing the above three tables for HIV-1 similarity analysis gives Table 6.9.

Table 6.10, Table 6.11 and Table 6.12 contain information from analyzing the similarity to the human proteome for HIV-2 at the 5-mer, 6-mer and 7-mer level respectively. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. Summarizing the above three tables for HIV-2 similarity analysis gives Table 6.13.

Table 6.5: Summary of chi-square analysis for each viral proteome (HIV-1, HIV-2 and Influenza A virus) at the 5-, 6- and 7-mer level. Two different techniques are conducted for each viral proteome.

k-**mers**	**HIV-1**					
	Technique 1			**Technique 2**		
	χ^2	**d.f.**	p-**value**	χ^2	**d.f.**	p-**value**
5-mer	62.9146	34	1.8000E-3	36.7741	15	1.4000E-3
6-mer	86.5532	35	2.9534E-6	49.0854	15	1.6970E-5
7-mer	33.4024	16	6.5000E-3	14.7212	15	4.7170E-1

k-**mers**	**HIV-2**					
	Technique 1			**Technique 2**		
	χ^2	**d.f.**	p-**value**	χ^2	**d.f.**	p-**value**
5-mer	86.9650	36	4.1886E-6	30.1165	15	1.1500E-2
6-mer	110.9699	36	1.4367E-9	44.7155	15	8.4902E-5
7-mer	72.8199	24	8.1132E-7	55.7509	15	1.3349E-6

k-**mers**	**Influenza A virus**					
	Technique 1			**Technique 2**		
	χ^2	**d.f.**	p-**value**	χ^2	**d.f.**	p-**value**
5-mer	87.1936	43	7.7963E-5	23.1269	9	5.9000E-3
6-mer	72.4257	43	3.3000E-3	19.4674	9	2.1500E-2
7-mer	31.6968	16	1.0900E-2	19.4987	9	2.1300E-2

Table 6.6: Summary information from analyzing the similarity to the human proteome for HIV-1 at the 5-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 9.505$ for $1 \leq i \leq 34; E_{35} = 12.8317$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**HIV-1 proteins**	**Conclusion**
0001-0100	2	5.925831142	Gag poly (p17)	humanlike
0301-0400	18	7.592322462	Gag poly (p24)	non-humanlike
0501-0600	3	4.451870068	Gag-Pol (p17)	humanlike
0801-0900	17	5.910049974	Gag-Pol (p24,p2,p7,p1,p6)	non-humanlike
1901-2000	16	4.438193056	Gag-Pol (Integrase), Env gp160	non-humanlike
2401-2500	3	4.451870068	Env gp160	humanlike
2801-2900	3	4.451870068	Nef	humanlike

Position	N **(**$ExpectedN$**)**	$(E_i - O_i)^2/E_i$	**HIV-1 proteins**	**Conclusion**
0001-0131	4 (12.451)	5.736037346	Gag poly (p17)	humanlike
0508-0638	4 (12.451)	5.736037346	Gag-Pol (p17)	humanlike
0870-1006	21 (13.022)	4.887765627	Gag-Pol (p2,p7,p1,p6)	non-humanlike
3162-3273	1 (10.646)	8.739931993	Rev	humanlike
3274-3461	27 (17.869)	4.665910851	Vif	non-humanlike

Table 6.7: Summary information from analyzing the similarity to the human proteome for HIV-1 at the 6-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 29.75$ for $1 \leq i \leq 35; E_{36} = 7.735$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**HIV-1 proteins**	**Conclusion**
0001-0100	46	8.876050420	Gag poly (p17)	humanlike
0501-0600	42	5.044117647	Gag-Pol (p17)	humanlike
2001-2100	15	7.313025210	Env gp160	non-humanlike
2101-2200	18	4.640756303	Env gp160	non-humanlike
2301-2400	17	5.464285714	Env gp160	non-humanlike
2601-2700	46	8.876050420	Env gp160	humanlike
2701-2800	41	4.254201681	Env gp160	humanlike
2801-2900	42	5.044117647	Nef	humanlike
3101-3200	48	11.19537815	Vpu, Rev	humanlike

Position	N ($ExpectedN$)	$(E_i - O_i)^2/E_i$	**HIV-1 proteins**	**Conclusion**
0001-0131	58 (38.973)	9.289167603	Gag poly (p17)	humanlike
0507-0637	58 (38.973)	9.289167603	Gag-Pol (p17)	humanlike
1665-1947	59 (84.194)	7.538988954	Gag-Pol (Integrase)	non-humanlike
3156-3266	54 (33.023)	13.32509248	Rev	humanlike

Table 6.8: Summary information from analyzing the similarity to the human proteome for HIV-1 at the 7-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. It uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 6.37$ for $1 \leq i \leq 16; E_{17} = 10.1$.

Position	N	$(E_i - O_i)^2/E_i$	**HIV-1 proteins**	**Conclusion**
1601-1800	1	4.526985871	Gag-Pol (RT,Integrase)	non-humanlike
2601-2800	17	17.7389168	Env gp160	humanlike

Table 6.9: Summary of all the conclusions of regions/proteins from HIV-1 proteome being humanlike/non-humanlike at the 5, 6, 7-mer levels.

k-mers	Technique 1		Technique 2	
	humanlike	**non-humanlike**	**humanlike**	**non-humanlike**
5-mer	Gag poly (p17) Gag-Pol (p17) Env gp160 Nef	Gag poly (p24) Gag-Pol (p24) Env gp160 Gag-Pol(Integrase) Gag-Pol (p2,p7,p1,p6)	Gag poly (p17) Gag-Pol (p17) Rev	Gag-Pol (p2,p7,p1,p6) Vif
6-mer	Gag poly (p17) Gag-Pol (p17) Env gp160 Nef Vpu Rev	Env gp160	Gag poly (p17) Gag-Pol (p17) Rev	Gag-Pol (Integrase)
7-mer	Env gp160	Gag-Pol (RT) Gag-Pol (Integrase)		

Table 6.10: Summary information from analyzing the similarity to the human proteome for HIV-2 at the 5-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 10.26$. Note that the last 23 k-mers are not included since there are no non-overlaps beyond position 3700. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**HIV-2 proteins**	**Conclusion**
0201-0300	21	11.24245614	Env gp160	non-humanlike
0301-0400	25	21.17617934	Env gp160	non-humanlike
0501-0600	2	6.649863548	Env gp160	humanlike
1901-2000	3	5.137192982	Gag-Pol (RT)	humanlike
2901-3000	17	4.427641326	Vif	non-humanlike

Position	N **(**$ExpectedN$**)**	$(E_i - O_i)^2/E_i$	**HIV-2 proteins**	**Conclusion**
1367-1465	2 (10.158)	6.551778303	Gag-Pol (Protease)	humanlike
2831-3041	33 (21.65)	5.950230947	Vif	non-humanlike

Table 6.11: Summary information from analyzing the similarity to the human proteome for HIV-2 at the 6-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 30.64$ for $1 \leq i \leq 36; E_{37} = 34.9296$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**HIV-2 proteins**	**Conclusion**
0201-0300	11	12.58908616	Env gp160	non-humanlike
0301-0400	16	6.995091384	Env gp160	non-humanlike
0501-0600	48	9.835822454	Env gp160	humanlike
1101-1200	12	11.33973890	Gag-Pol (p24)	non-humanlike
1201-1300	45	6.730078329	Gag-Pol (p24,p2,p7,p1,p6)	humanlike
1301-1400	42	4.211801567	Gag-Pol (p2,p7,p1,p6,Protease)	humanlike
2001-2100	16	6.995091384	Gag-Pol (RT,Integrase)	non-humanlike
2301-2400	44	5.825378590	Gag-Pol (Integrase), Gag poly (p17)	humanlike
2801-2900	17	6.072114883	Gag poly (p2,p7,p1,p6), Vif	non-humanlike
3001-3100	46	7.700052219	Vif, Nef	humanlike
3501-3600	44	5.825378590	Rev, Vpr	humanlike

Position	N ($ExpectedN$)	$(E_i - O_i)^2/E_i$	**HIV-2 proteins**	**Conclusion**
0854-0987	55 (41.059)	4.733468448	Gag-Pol (p17)	humanlike
1218-1365	64 (45.348)	7.671718797	Gag-Pol (p2,p7,p1,p6)	humanlike
2024-2311	66 (88.246)	5.608010743	Gag-Pol (Integrase)	non-humanlike
2312-2445	55 (41.059)	4.733468448	Gag poly (p17)	humanlike
3508-3607	43 (30.641)	4.984983551	Vpr	humanlike

Table 6.12: Summary information from analyzing the similarity to the human proteome for HIV-2 at the 7-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 5.587$ for $1 \leq i \leq 24$; $E_{25} = 3.9109$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**HIV-2 proteins**	**Conclusion**
0451-0600	11	5.244419008	Env gp160	humanlike
0601-0750	0	5.587000000	Env gp160	non-humanlike
0751-0900	11	5.244419008	Env gp160, Gag-Pol (p17)	humanlike
1201-1350	17	23.31422391	Gag-Pol (p24,p2,p7,p1,p6)	humanlike
3601-3705	11	12.85007001	Vpx	humanlike
Position	N **(**$ExpectedN$**)**	$(E_i - O_i)^2/E_i$	**HIV-2 proteins**	**Conclusion**
1217-1364	16 (5.513)	19.94869744	Gag-Pol (p2,p7,p1,p6)	humanlike
2674-2824	11 (5.624)	5.138935989	Gag poly (p2,p7,p1,p6)	humanlike
3600-3705	11 (3.948)	12.59642958	Vpx	humanlike

Table 6.13: Summary of all the conclusions of regions/proteins from HIV-2 proteome being humanlike/non-humanlike at the 5, 6, 7-mer levels.

k**-mers**	**Technique 1**		**Technique 2**	
	humanlike	**non-humanlike**	**humanlike**	**non-humanlike**
5-mer	Env gp160 Gag-Pol (RT)	Env gp160 Vif	Gag-Pol (Protease)	Vif
6-mer	Env gp160 Gag-Pol (p17) Gag-Pol (p2,p7,p1,p6) Gag-Pol (Protease) Nef Rev, Vpr	Env gp160 Gag-Pol (p24) Gag poly (p2,p7,p1,p6) Vif	Gag-Pol (p17) Gag-Pol (p2,p7,p1,p6)	Gag-Pol(Integrase)
7-mer	Env gp160 Gag-Pol (p17) Gag-Pol (p2,p7,p1,p6) Vpx	Env gp160	Gag-Pol (p2,p7,p1,p6) Gag poly (p2,p7,p1,p6) Vpx	

Table 6.14, Table 6.15 and Table 6.16 contain information from analyzing the similarity to the human proteome for Influenza A virus at the 5-mer, 6-mer and 7-mer level respectively. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. Summarizing the above three tables for Influenza A virus similarity analysis gives Table 6.17.

Table 6.14: Summary information from analyzing the similarity to the human proteome for Influenza A virus at the 5-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 8.49$ for $1 \leq i \leq 43; E_{44} = 10.78$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**Influenza A virus proteins**	**Conclusion**
0401-0500	2	4.961142521	Matrix 1	humanlike
0801-0900	18	10.65254417	Hemagglutinin	non-humanlike
1501-1600	15	4.991766784	Neuraminidase	non-humanlike
1601-1700	20	15.60425206	Neuraminidase	non-humanlike
2301-2400	2	4.961142521	Polymerse (gene: None)	humanlike
3301-3400	15	4.991766784	Polymerse (gene: PB1)	non-humanlike
4301-4427	1	8.872764378	Polymerse (gene: PB2)	humanlike
Position	N **(**$ExpectedN$**)**	$(E_i - O_i)^2/E_i$	**Influenza A virus proteins**	**Conclusion**
0344-0591	9 (21.063)	6.908606039	Matrix 1	humanlike
1249-1713	58 (39.494)	8.671495316	Neuraminidase	non-humanlike

Self-similarity

A chi-square analysis on the distribution of overlapping/non-overlapping k-mers within human protein Dsg3 are performed using two different techniques of dividing into segments. In the case of 5-mers, non-overlapping 5-mers are considered since there exists more overlaps than non-overlaps at the 5-mer level. However, with 6-mers and 7-mers, there are more non-overlapping k-mers than overlapping ones. Therefore, it makes more sense to look at the distribution of the locations of 6-mers and 7-mers which also occur in the remaining human proteins. Table 6.18 contains summarized information of the calculated χ^2 value, degree of freedom (d.f.) and p-value for each analysis. By conventional criteria, the difference of analysis at the 5-mer level is not considered statistically significant. However, the differences of analysis at the 6-mer and 7-mer level are considered statistically significant. In summary, there is no structure (clustering) to the regions of self-similarity at the 5-mer level while there is clustering at the 6-mer and 7-mer level.

Table 6.15: Summary information from analyzing the similarity to the human proteome for Influenza A virus at the 6-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 27.055$ for $1 \leq i \leq 43$; $E_{44} = 31.654$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**Influenza A virus proteins**	**Conclusion**
0901-1000	14	6.299501940	Hemagglutinin	non-humanlike
1501-1600	14	6.299501940	Neuraminidase	non-humanlike
2801-2900	39	5.273813528	Polymerse (gene: None)	humanlike
3201-3300	16	4.517206616	Polymerse (gene: PB1)	non-humanlike
3701-3800	13	7.301534836	Polymerse (gene: PB2)	non-humanlike
3901-4000	38	4.427759194	Polymerse (gene: PB2)	humanlike
4301-4417	46	6.506240126	Polymerse (gene: PB2)	humanlike

Position	N **(**$ExpectedN$**)**	$(E_i - O_i)^2/E_i$	**Influenza A virus proteins**	**Conclusion**
0342-0588	85 (66.825)	4.943219229	Matrix 1	humanlike
2912-3663	172 (203.45)	4.861649054	Polymerse (gene: PB1)	non-humanlike

Table 6.16: Summary information from analyzing the similarity to the human proteome for Influenza A virus at the 7-mer level. Only observations that are different from random with statistical significance have been listed and conclusions of being humanlike/non-humanlike are drawn accordingly. First table uses the technique of dividing the viral proteome into evenly-sized segments: $E_i = 5.446$ for $1 \leq i \leq 16$; $E_{17} = 8.866$. The second table uses the technique of dividing the viral proteome into individual protein segments: the expected matches vary in proportion to the proteins' length.

Position	N	$(E_i - O_i)^2/E_i$	**Influenza A virus proteins**	**Conclusion**
0001-0250	12	7.887424899	Nonstructural 1, Nonstrucrual 2	humanlike
2001-2250	11	5.664141755	Nucleocapsid, Polymerse (gene: None)	humanlike
3501-3750	0	5.446000000	Polymerse (gene: PB1, gene: PB2)	non-humanlike

Position	N **(**$ExpectedN$**)**	$(E_i - O_i)^2/E_i$	**Influenza A virus proteins**	**Conclusion**
0001-0224	10 (4.88)	5.371803279	Nonstructural 1	humanlike
2904-3654	5 (16.359)	7.887210771	Polymerse (gene: PB1)	non-humanlike

Table 6.17: Summary of all the conclusions of regions/proteins from Influenza A virus proteome being humanlike/non-humanlike at the 5, 6, 7-mer levels.

k-mers	Technique 1		Technique 2	
	humanlike	**non-humanlike**	**humanlike**	**non-humanlike**
5-mer	Matrix 1 Polymerse(gene:None) Polymerse(gene:PB2)	Hemagglutinin Neuraminidase Polymerse(gene:PB1)	Matrix 1	Neuraminidase
6-mer	Polymerse(gene:None) Polymersegene: PB2)	Hemagglutinin Neuraminidase Polymerse(gene:PB1) Polymerse(gene:PB2)	Matrix 1	Polymerse(gene:PB1)
7-mer	Nonstructural 1 Nonstructural 2 Nucleocapsid Polymerse(gene:None)	Polymerse(gene:PB1) Polymerse(gene:PB2)	Nonstructural 1	Polymerse(gene:PB1)

Table 6.18: Summary of chi-square analysis for human protein Dsg3 at the 5-, 6- and 7-mer level.

k-mers	Desmoglein 3		
	χ^2	**d.f.**	p-**value**
5-mer	4.0710	6	6.6710E-1
6-mer	17.1571	9	4.6300E-2
7-mer	45.7352	9	6.7434E-7

6.2.4 Discussion

Given that the overlapping k-mers of the HIV-1 and HIV-2 proteomes are not distributed at random by comparing the χ^2 value with the values in a standard chi-square table, each segment is analyzed to indicate whether or not being "humanlike" or "non-humanlike". Table 6.9 and Table 6.13 give the summarized results of the similarity analysis for HIV-1 proteome. It shows that portions of Env gp160 protein being humanlike and the other portions of it being non-humanlike. Env gp160 contains two proteins: surface protein gp120 (SU) and trans-membrane protein gp41 (TM). The surface protein attaches the virus to the host lymphoid cell by binding to the primary receptor CD4. This interaction induces a structural rearrangement creating a high affinity binding site. Thus, the surface protein being humanlike agrees with that it is not easy to be recognized by MHC molecules and induces immune response. However, the other protein, transmembrane protein gp41, acts as a class I viral fusion protein. Membranes fusion leads to delivery of the nucleocapsid into the cytoplasm. As we know, nucleocapsid consists of a core of nucleic acid enclosed in a protein coat, it protects the key information of the virus. The trans-membrane protein being non-humanlike agree with that it can easily destroy the cell membrane and deliver the nucleocapsid into the cytoplasm.

Table 6.17 are also investigated similarly to give indications whether certain segments/regions are being humanlike/non-humanlike. It shows that Matrix protein 1, Polymerse (gene: None), Polymerse (gene: PB2) and Nonstructural protein 1 are humanlike while Hemagglutinin portein, Neuraminidase protein and Polymerse (gene: PB1) are non-humanlike. Matrix protein 1 plays critical roles in virus replication, from virus entry and uncoating to assembly and budding of the virus particle. It forms a continuous shell on the inner side of the layer where it binds the ribonucleocapsids. Therefore, Matrix protein 1 being non-humanlike agrees with its major function of virus replication. In addition, Hemagglutinin protein, which is a Class I viral fusion protein, is non-humanlike as well. It binds to sialic acid-containing receptors on the cell surface, bringing about the attachment of the virus particle to the cell. It plays a major role in the determination of host range restriction and virulence. Therefore, Hemagglutinin protein being non-humanlike also match with its biological function that it destroy the cell surface and help with the delivery of virus particle.

More results from Table 6.9, Table 6.13 and Table 6.17 need to be validated if characteristics of certain viral proteins are known. Such exploration needs collaborations with professionals from molecule biology area.

CHAPTER 7

SUMMARY AND FUTURE WORK

7.1 Summary and Discussion

In this research work, we explored the theme of whether the regions of immunological significance come from the regions of low host-similarity as well as the theme of whether the locality clustering of overlapping occurrences in the viral proteome is statistically significant. The purpose of the thesis is a "survey approach" to questions related to immunology, especially in the area of discriminating immunological self from non-self. Similar work is also explored by other researchers. One scientist studied the question of whether the peptides of nine amino acids (9-mers) that are typically used in MHC class I presentation are sufficiently unique for self:non-self discrimination. Her results show that the 9-mers used in MHC class I presentation tend to carry sufficient information to detect non-self peptides amongst self peptides [3]. By enumerating distinct 9-mers for a variety of microorganisms, she found that the probability that a foreign peptide also occurs in the human self is about 0.2%. A small overlap between self and nonself makes sense for several reasons. First, if more foreign epitope that overlaps with self, the chances of autoimmunity will increase. Second, if we have smaller overlap between self and nonself, more peptides will remain as potential targets for detecting the presence of the pathogen. Another scientist studied the targets of the immune response in autoimmune diseases by applying the principle of nonself-discrimination in the identification of the autoimmunogenic peptide sequences. Her results show that low level of sequence similarity to the host's proteome may modulate peptide epitopicity [16]. It motivates the themes that are investigated in my thesis and provides some validations of results from wet lab experiments.

7.2 Future Work

7.2.1 Peptide Universes and Corresponding Overlaps

If the proteome of each organism is a single, linear sequence, there is no problem when generating peptide universe for an organism or the overlaps. This is already the case for viruses which have a polyprotein. However, for other organisms, their (multiple) protein sequences result in a procedural problem. The proteins in the proteome can be linearized into a single, long sequence of amino acids easily enough: if the

protein sequences are stored in a multi-FASTA file, the sequence header information for all but the first sequence is deleted. Unfortunately, a problem arises due to the linearization. Artificial peptides may be produced which may affect the results of the clustering analysis. These artificial peptides come from subsequences consisting of the end of one protein and the beginning of the following protein. For example, *Salmonella typhi* (Taxonomic ID:601) contains 4716 proteins. The first protein sequence, with identifier Q56114, ends with amino acids "LCEAIVAVL" and the second protein sequence, with identifier P40674, starts with amino acids "SLNFLDFEQ". Linearizing the protein sequences of *Salmonella typhi* creates the artificial amino acid subsequence "LCEAIVAVLSLNFLDFEQ" as part of the overall sequence. When the k-mers of *Salmonella typhi* are produced, peptides from this artificial amino acid subsequence are also produced. An improved technique needs to be proposed to circumvent the "artificial peptide" problem described above.

7.2.2 Filtering of Ambiguities

As discussed in Section 4.1, all three filtering strategies have advantages as well as disadvantages. The first method causes errors of mismatch, which will make the counting of overlaps between two organisms incorrect, and errors of missing information. The other two methods lose information when performing the filtering. Although it shows that the computational cost of the third strategy is within our set threshold while maintaining the minimum information loss, it is possible that we could try possible matches and leave the ambiguities in the proteome file.

7.2.3 Expected Number of Overlaps

In Section 5.2.1, the expression of calculating the expected number of overlaps is developed. The upper bound is used to calculate the expected unique overlaps and thus this expected number is mostly higher than one should be. Also, random sampling is assumed when deriving the equation for calculating the expected number of overlapping occurrences including duplicates. Therefore, this expected number is mostly lower than one should be. In order to make these expected numbers be more statistically accurate, a more refined analysis would take into account the dependence caused by the overlap of neighboring protein sequences and the relative frequency of occurrence of individual amino acids. That is, the amino acids that tend to appear in human proteome are more likely to appear in viral proteomes as well. The dependency between two samples could cause the expected number of overlapping occurrences including duplicates become lower than one should be. In further stage of the work, the equations for calculating the expected number of overlaps could be refined and reformulated.

7.2.4 Proteomic Similarity Analysis

In chapter 5, the host-similarity between viral proteom (HIV-1) and human proteome is investigated and it turns out that there is no correlation between the regions of low host-similarity and immunodominants. One possible reason is that there is too few overlaps between HIV-1 and human at the 7-mer level. Thus, overlaps at the 5-mer and 6-mer levels could be investigated in the future. As introduced in Chapter 2, the typical length of peptides presented to $CD8^+$ T cells by MHC-I molecules is 8 - 13 residues while to B cells is 5 - 6 residues. Therefore, to perform the similarity analysis at the 5-mer and 6-mer level, the list of all HIV antibody binding sites mapped to within a region of 21 amino acids or less should be used to indicate regions of immunological significance.

Further, more human viral proteomes can be explored such as Influenza A virus, HCV, SARS etc. Since we need to compare the regions of low host-similarity with regions of immunological significance, completely and well-studied viruses are good candidates.

7.2.5 Locality Clustering Analysis

In an organism's proteome, there typically exist regions such as amino sequences of "LLLLLLLLL". These are called "low complexity regions". Low complexity regions are portions of protein sequence of biased composition including homo-polymeric runs, short-period repeats, and more subtle over-representations of one or a few residues. These low complexity regions appear frequently in proteome data. Therefore, they may cause violation of some of the assumptions made when looking for statistically significant sequences. They should be eliminated from the similarity analysis for this reason. Thus in the future, a filtering technique may be introduced for this purpose.

In chapter 6, structure (clustering) to the regions of host-similarity is investigated. Only three viral proteomes are investigated at the 5-, 6- and 7-mer levels. No analysis is done for 8- and 9-mer levels since there do not exist enough overlaps between viral and human proteomes. In future, more human affected viruses could be analyzed, even well-studied bacteria. As mentioned earlier, bacteria usually have more overlaps with human at the 8- and 9-mer levels. Therefore, such clustering analysis could be done for bacteria at the 9-mer level. Also, for self-similarity analysis, more human proteins which are known to be involved in autoimmune diseases could be explored in the future.

7.2.6 Phylogenetic Analysis

The degree of overlap (similarity) between the generated peptide universes could be used for building phylogenetic trees especially relevant to the immunological context. It would be interesting if the resultant trees differed from trees generated by more usual techniques (multi-sequence alignment of genomic or protein sequences). In the future, such an investigation will be considered.

REFERENCES

[1] Mik Bickis. Personal communication, Department of Mathematics and Statistics, 2006.

[2] P. W. Bryant, A. M. Lennon-Dumenil, E. Fiebiger, C. Lagaudriere-Gesbert, and H. L. Ploegh. Proteolysis and antigen presentation by mhc class ii molecules. *Adv Immunol*, 80:71 – 80, 2002.

[3] Nigel J. Burroughs, Rob J. Boer, and Can Kesmir. Discriminating self from nonself with short peptides from large proteoms. *Immunogenetics*, 10:1 – 10, 2004.

[4] Richard Coico, Geoffrey Sunshine, and Eli Benjamini. *Immunology : a short course*. Wiley-Liss, Canada, 5th edition, 2003.

[5] J. L. Cornette, H. Margalit, J. A. Berzofsky, and C. Delisi. Periodic variation in side-chain polarities of t-cell antigenic peptides correlates with their structure and activity. *Proc Natl Acad Sci*, 92:8368 – 8372, USA 2002.

[6] Gusfield Dan. *Algorithms on Strings, Trees, and Sequences*. Cambridge, USA, 1st edition, 1997.

[7] V. H. Engelhard. Structure of peptides associated with class I and class II MHC molecules. *Annu Rev Immunol*, 12:181 – 207, 1994.

[8] D. T. Fearon and R. M. Locksley. The instructive role of innate immunity in the acquired immune response. *Science*, 272:50 – 60, 1996.

[9] R. N. Germain and D. H. Margulies. The biochemistry and cell biology of antigen processing and presentation. *Annu Rev Immunol*, 11:403 – 450, 1993.

[10] Baum H., Davies H., and Peakman M. Database screening for molecular mimicry. *Immunology Today*, 18(5):252 – 253, 1997.

[11] Charles A. Janeway, Paul Travers, Mark Walport, and Mark J. Shlomchik. *Immuno Biology*. Garland Science, New York and London, 6th edition, 2005.

[12] Thomas J. Kindt, Richard A. Goldsby, and Barbara A. Osborne. *Immunology*. Freeman, USA, 6th edition, 1992.

[13] Tung K.S. Mechnism of self-tolerence and events leading to autoimmune disease and autoantibody response. *Clinical Immunology*, 73:275 – 282, 1994.

[14] C. Kuttler, A. K. Nussbaum, T. P. Dick, H. G. Rammensee, H. Schild, and K. P. Hadeler. An algorithm for the prediction of proteasomal cleavages. *J. Mol Biol*, 298:417 – 429, 2000.

[15] MS Lim and KSJ Elenitoba-Johnson. Proteomics in pathology research. *Lab Invest*, 84:1227 – 1244, 2004.

[16] A. Lucchese, A. Mittelman, M. S. Lin, D. Kanduc, and A. A. Sinha. Epitope definition by proteomic similarity analysis: identification of the linear determinant of the anti-Dsg3 MAb 5H10. *J Trans Med*, 2(1):43 – 50, 2004.

[17] G. E. Meister, C. G. P. Roberts, J. A. Berzofsky, and A. S. De Groot. Two novel T cell epitope prediction algorithms based on MHC-binding motifs; comparison of predicted and published epitopes from Mycobacterium tuberculosis and HIV protein sequences. *Vaccine*, 13(6):581 – 591, 1995.

[18] A. Mittelman, A. Lucchese, A. A. Sinha, and D. Kanduc. Monoclonal and polyclonal humoral immune response to EC HER-2/NEU peptides with low similarity to the host's proteome. *Int J Cancer*, 98:741 – 747, 2002.

[19] A. Mittelman, A. Lucchese, A. A. Sinha, and D. Kanduc. Monoclonal and polyclonal humoral immune response to EC HER-2/NEU peptides with low similarity to the host's proteome. *Int J. Cancer*, 98:741 – 747, 2002.

[20] C. Natale, T. Giannini, A. Lucchese, and D. Kanduc. Computer-assisted analysis of molecular mimicry between HPV16 E7 oncoprotein and human protein seqeunces. *Immunol Cell Biol*, 78:580 – 585, 2000.

[21] C. Natale, T. Giannini, A. Lucchese, and D. Kanduc. Computer-assisted analysis of molecular mimicry between HPV16 E7 oncoprotein and human protein sequences. *Immunol Cell Biol*, 78:580 – 585, 2000.

[22] Jacqueline Parkin and Bryony Cohen. An overview of the immune system. *Lancet*, 357:1777 – 1789, 2001.

[23] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanovic. Syfpeithi: database for MHC ligands and peptide motifs. *Immunogenetics*, 50:213 – 219, 1999.

[24] E. J. Reits, C. Neijssen, W. Herberts, W. Benckhuijsen, L. Janssen, J. W. Drijfhout, and J. Neefjes. A Major Role for TPPII in Trimming Proteasomal Degradation Products for HMC Class I Antigen Presentation. *Immunity*, 20:495 – 504, 2004.

[25] L. Stoltze, M. Schirle, G. Schwarz, M. W. Thompson, L. B. Hersh, S. Kalbacher, H. Stevanovic, H. G. Rammensee, and H. Schild. Two new proteases in the MHC class I processing pathway. *Nat Immunol*, 1:413 – 420, 2000.

[26] C. B. Thompson. New insights into V(D)J recombination and its role in the evolution of the immune system. *Immunity*, 3:331 – 340, 1995.

[27] J. Willers, A. Lucchese, D. Kanduc, and S. Ferrone. Molecular mimicry of phage displayed peptides mimicking GD3 ganglioside. *Peptides*, 20(9):1021 – 1026, 1999.

[28] J. Willers, A. Lucchese, D. Kanduc, and S. Ferrone. Nikecykar mimicry of phage displayed peptides mimicking gd3 ganglioside. *Peptides*, 20:1021 – 1026, 1999.

[29] J. Yewdell, L. C. Anton, I. Bacik, U. Schubert, H. L. Snyder, and J. R. Bennink. Generating MHC class I ligands from viral gene products. *Immunol Rev*, 172:97 – 110, 1999.

APPENDIX A

MODEL ORGANISM PROTEOME DESCRIPTIONS

601.FASTA 4716 proteins from *Salmonella typhi* downloaded from
http://www.ebi.ac.uk/proteome/index.html on June 29, 2004

3702.FASTA 26150 proteins from *Arabidopsis thaliana* downloaded from
http://www.ebi.ac.uk/integr8 on November 7, 2004

6239.FASTA 21821 proteins for *Caenorhabditis elegans* downloaded from
http://www.ebi.ac.uk/proteome/index.html on June 22, 2004

7227.FASTA 19964 proteins from *Drosophila melanogaster* downloaded from
http://www.ebi.ac.uk/integr8 on November 7, 2004

9606.FASTA 34044 proteins for *Homo sapiens* downloaded from
http://www.ebi.ac.uk/integr8 on November 11, 2004

10090.FASTA 29051 proteins from *Mus musculus* downloaded from
http://www.ebi.ac.uk/integr8 on November 7, 2004

10116.FASTA 5863 proteins from *Rattus norvegicus* downloaded from
http://www.ebi.ac.uk/integr8 on July 9, 2004

36329.FASTA 5250 proteins from *Plasmodium falciparum* downloaded from
http://www.ebi.ac.uk/integr8 on November 7, 2004

50339.FASTA 1524 proteins from *Thermoplasma volcanium* (Accession number BA000011)
downloaded from http://www.ebi.ac.uk/integr8 on November 14, 2004

85962.FASTA 1555 proteins from *Helicobacter pylori* (*Campylobacter pylori*)
(Accession number AE000511) downloaded from http://www.ebi.ac.uk/integr8 on November 14, 2004

115711.FASTA 1110 proteins from *Chlamydia pneumoniae* strain AR39
(Accession number AE002161) downloaded from
http://www.ebi.ac.uk/integr8 on November 14, 2004

224308.FASTA 4105 proteins from *Bacillus subtilis* (Accession number AL009126)
downloaded from http://www.ebi.ac.uk/integr8 on November 14, 2004

243277.FASTA 3785 proteins from *Vibrio cholerae* downloaded from
http://www.ebi.ac.uk/integr8 on November 14, 2004

APPENDIX B

SUMMARY INFORMATION FOR 13 MODEL ORGANISMS

This appendix contains 3 tables giving summary information for the 13 model organisms at the 9-mer level, one for each filtering strategy.

Table B.1: Summary information of the 13 model organisms for the filtering strategy of including all proteins, irrespective of whether they contain ambiguous amino acids (B, X or Z).

Taxonomic ID	Name	# of proteins	# of unique 9-mers	# of 9-mer occurrences
601	Salmonella typhi	4716	1353971	1369302
3702	Arabidopsis thaliana	26150	9805060	11127054
6239	Caenorhabditis elegans	21821	7808847	9360910
10116	Rattus norvegicus	5863	2573426	2827972
36329	Plasmodium falciparum	5250	3738541	3948101
50339	Thermoplasma volcanium	1524	437512	440589
85962	Helicobacter pylori	1555	473803	479483
115711	Chlamydia pneumoniae	1110	353233	354710
224308	Bacillus subtilis	4105	1179278	1189168
7227	drosophila melanogaster	19964	7217497	10741502
243277	Vibrio cholerae	3785	1115331	1124273
9606	Homo sapiens	34044	10744680	15367806
10090	Mus musculus	29051	9503051	12842290

Table B.2: Summary information of the 13 model organisms for the filtering strategy of removing any proteins that contain any ambiguous amino acids (B, X or Z).

Taxonomic ID	Name	# of proteins	# of unique 9-mers	# of 9-mer occurrences
601	Salmonella typhi	4714	1353169	1368473
3702	Arabidopsis thaliana	26145	9804087	11124178
6239	Caenorhabditis elegans	21809	7806071	9356423
10116	Rattus norvegicus	5786	2542425	2793097
				Continued…

Table B.2 – continued from previous page				
Taxonomic ID	**Name**	**# of proteins**	**# of unique 9-mers**	**# of 9-mer occurrences**
36329	Plasmodium falciparum	5245	3735883	3945141
50339	Thermoplasma volcanium	1523	437176	440200
85962	Helicobacter pylori	1544	469880	474271
115711	Chlamydia pneumoniae	1093	344862	346245
224308	Bacillus subtilis	4103	1178837	1188706
7227	drosophila melanogaster	19926	7208445	10709861
243277	Vibrio cholerae	3772	1106020	1114925
9606	Homo sapiens	33800	10670465	15224726
10090	Mus musculus	28816	9444529	12721665

Table B.3: Summary information of the 13 model organisms for the filtering strategy of removing only generated 9-mers which contain ambiguous amino acids (B, X or Z).

Taxonomic ID	**Name**	**# of proteins**	**# of unique 9-mers**	**# of 9-mer occurrences**
601	Salmonella typhi	4716	1353953	1369284
3702	Arabidopsis thaliana	26150	9805006	11127000
6239	Caenorhabditis elegans	21821	7808774	9360801
10116	Rattus norvegicus	5863	2572460	2826217
36329	Plasmodium falciparum	5250	3738467	3947981
50339	Thermoplasma volcanium	1524	437503	440580
85962	Helicobacter pylori	1555	473682	479362
115711	Chlamydia pneumoniae	1110	353071	354548
224308	Bacillus subtilis	4105	1179260	1189150
7227	drosophila melanogaster	19964	7216967	10740972
243277	Vibrio cholerae	3785	1115183	1124125
9606	Homo sapiens	34044	10739022	15359070
10090	Mus musculus	29051	9399769	12838611

APPENDIX C

OVERLAP TABLES FOR 13 MODEL ORGANISMS

The following table is the comparison of 13×13 pairs of model organisms by using the first filtering strategy. The technique of counting the unique 9-mers in the overlap of both proteomes is used here. See next page for the complete table.

Proteome Name	**Taxonomic ID**	601		3702		6239		10116		36329		50339		85962		115711		224308		7227		243277		9606		10090	
Taxonomic ID	**# of unique 9-mers**	1353971		9805060		7808847		2573426		3738541		437512		473803		353233		1179278		7217497		1115331		10744680		9503051	
S. typhi				3839	3.915E-04	1717	2.199E-04	1073	4.170E-04	894	2.391E-04	536	1.225E-03	2496	5.268E-03	1452	4.111E-03	5409	4.587E-03	2235	3.097E-04	50474	4.525E-02	2613	2.432E-04	2136	2.248E-04
601	1353971			2.835E-03	1.110E-06	1.268E-03	2.788E-07	7.925E-04	3.304E-07	6.603E-04	1.579E-07	3.959E-04	4.850E-07	1.843E-03	9.711E-06	1.072E-03	4.408E-06	3.995E-03	1.832E-05	1.651E-03	5.112E-07	3.728E-02	1.687E-03	1.930E-03	4.693E-07	1.578E-03	3.546E-07
A. thaliana		3839	2.835E-03			23329	2.988E-03	14200	5.518E-03	12930	3.459E-03	779	1.781E-03	1532	3.233E-03	1383	3.915E-03	3422	2.902E-03	30158	4.178E-03	3347	3.001E-03	35912	3.342E-03	34201	3.599E-03
3702	9805060	3.915E-04	1.110E-06			2.379E-03	7.108E-06	1.448E-03	7.991E-06	1.319E-03	4.561E-06	7.945E-05	1.415E-07	1.562E-04	5.052E-07	1.410E-04	5.522E-07	3.490E-04	1.013E-06	3.076E-03	1.285E-05	3.414E-04	1.024E-06	3.663E-03	1.224E-05	3.488E-03	1.255E-05
C. elegans		1717	1.268E-03	23329	2.379E-03			34388	1.336E-02	9798	2.621E-03	530	1.211E-03	622	1.313E-03	533	1.509E-03	1418	1.202E-03	63286	8.768E-03	1653	1.482E-03	68134	6.341E-03	65690	6.913E-03
6239	7808847	2.199E-04	2.788E-07	2.988E-03	7.108E-06			4.404E-03	5.885E-05	1.255E-03	3.288E-06	6.787E-05	8.222E-08	7.965E-05	1.046E-07	6.826E-05	1.030E-07	1.816E-04	2.183E-07	8.104E-03	7.106E-05	2.117E-04	3.137E-07	8.725E-03	5.533E-05	8.412E-03	5.815E-05
R. norvegicus		1073	7.925E-04	14200	1.448E-03	34388	4.404E-03			5856	1.566E-03	274	6.263E-04	393	8.295E-04	347	9.824E-04	985	8.353E-04	55280	7.659E-03	922	8.267E-04	1224571	1.140E-01	1661771	1.749E-01
10116	2573426	4.170E-04	3.304E-07	5.518E-03	7.991E-06	1.336E-02	5.885E-05			2.276E-03	3.564E-06	1.065E-04	6.668E-08	1.527E-04	1.267E-07	1.348E-04	1.325E-07	3.828E-04	3.197E-07	2.148E-02	1.645E-04	3.583E-04	2.962E-07	4.759E-01	5.423E-02	6.457E-01	1.129E-01
P. falciparum		894	6.603E-04	12930	1.319E-03	9798	1.255E-03	5856	2.276E-03			344	7.863E-04	510	1.076E-03	419	1.186E-03	823	6.979E-04	11651	1.614E-03	845	7.576E-04	12072	1.124E-03	11749	1.236E-03
36329	3738541	2.391E-04	1.579E-07	3.459E-03	4.561E-06	2.621E-03	3.288E-06	1.566E-03	3.564E-06			9.201E-05	7.235E-08	1.364E-04	1.468E-07	1.121E-04	1.329E-07	2.201E-04	1.536E-07	3.116E-03	5.031E-06	2.260E-04	1.712E-07	3.229E-03	3.628E-06	3.143E-03	3.885E-06
T. volcani		536	3.959E-04	779	7.945E-05	530	6.787E-05	274	1.065E-04	344	9.201E-05			234	4.939E-04	136	3.850E-04	754	6.394E-04	518	7.177E-05	337	3.022E-04	629	5.854E-05	585	6.156E-05
50339	437512	1.225E-03	4.850E-07	1.781E-03	1.415E-07	1.211E-03	8.222E-08	6.263E-04	6.668E-08	7.863E-04	7.235E-08			5.348E-04	2.641E-07	3.108E-04	1.197E-07	1.723E-03	1.102E-06	1.184E-03	8.497E-08	7.703E-04	2.327E-07	1.438E-03	8.416E-08	1.337E-03	8.231E-08
H. pylori		2496	1.843E-03	1532	1.562E-04	622	7.965E-05	393	1.527E-04	510	1.364E-04	234	5.348E-04			879	2.488E-03	2033	1.724E-03	722	1.000E-04	2239	2.007E-03	843	7.846E-05	801	8.429E-05
85962	473803	5.268E-03	9.711E-06	3.233E-03	5.052E-07	1.313E-03	1.046E-07	8.295E-04	1.267E-07	1.076E-03	1.468E-07	4.939E-04	2.641E-07			1.855E-03	4.617E-06	4.291E-03	7.397E-06	1.524E-03	1.524E-07	4.726E-03	9.487E-06	1.779E-03	1.396E-07	1.691E-03	1.425E-07
C. pneumoniae		1452	1.072E-03	1383	1.410E-04	533	6.826E-05	347	1.348E-04	419	1.121E-04	136	3.108E-04	879	1.855E-03			1390	1.179E-03	661	9.158E-05	1289	1.156E-03	723	6.729E-05	705	7.419E-05
115711	353233	4.111E-03	4.408E-06	3.915E-03	5.522E-07	1.509E-03	1.030E-07	9.824E-04	1.325E-07	1.186E-03	1.329E-07	3.850E-04	1.197E-07	2.488E-03	4.617E-06			3.935E-03	4.638E-06	1.871E-03	1.714E-07	3.649E-03	4.217E-06	2.047E-03	1.377E-07	1.996E-03	1.481E-07
B. subtilis		5409	3.995E-03	3422	3.490E-04	1418	1.816E-04	985	3.828E-04	823	2.201E-04	754	1.723E-03	2033	4.291E-03	1390	3.935E-03			1583	2.193E-04	4460	3.999E-03	1908	1.776E-04	1765	1.857E-04
224308	1179278	4.587E-03	1.832E-05	2.902E-03	1.013E-06	1.202E-03	2.183E-07	8.353E-04	3.197E-07	6.979E-04	1.536E-07	6.394E-04	1.102E-06	1.724E-03	7.397E-06	1.179E-03	4.638E-06			1.342E-03	2.944E-07	3.782E-03	1.512E-05	1.618E-03	2.873E-07	1.497E-03	2.780E-07
D. melanogaster		2235	1.651E-03	30158	3.076E-03	63286	8.104E-03	55280	2.148E-02	11651	3.116E-03	518	1.184E-03	722	1.524E-03	661	1.871E-03	1583	1.342E-03			1686	1.512E-03	127941	1.191E-02	122322	1.287E-02
7227	7217497	3.097E-04	5.112E-07	4.178E-03	1.285E-05	8.768E-03	7.106E-05	7.659E-03	1.645E-04	1.614E-03	5.031E-06	7.177E-05	8.497E-08	1.000E-04	1.524E-07	9.158E-05	1.714E-07	2.193E-04	2.944E-07			2.336E-04	3.531E-07	1.773E-02	2.111E-04	1.695E-02	2.182E-04
V. cholerae		50474	3.728E-02	3347	3.414E-04	1653	2.117E-04	922	3.583E-04	845	2.260E-04	337	7.703E-04	2239	4.726E-03	1289	3.649E-03	4460	3.782E-03	1686	2.336E-04			1966	1.830E-04	1813	1.908E-04
243277	1115331	4.525E-02	1.687E-03	3.001E-03	1.024E-06	1.482E-03	3.137E-07	8.267E-04	2.962E-07	7.576E-04	1.712E-07	3.022E-04	2.327E-07	2.007E-03	9.487E-06	1.156E-03	4.217E-06	3.999E-03	1.512E-05	1.512E-03	3.531E-07			1.763E-03	3.225E-07	1.626E-03	3.101E-07
H. sapiens		2613	1.930E-03	35912	3.663E-03	68134	8.725E-03	1224571	4.759E-01	12072	3.229E-03	629	1.438E-03	843	1.779E-03	723	2.047E-03	1908	1.618E-03	127941	1.773E-02	1966	1.763E-03			3791399	3.990E-01
9606	10744680	2.432E-04	4.693E-07	3.342E-03	1.224E-05	6.341E-03	5.533E-05	1.140E-01	5.423E-02	1.124E-03	3.628E-06	5.854E-05	8.416E-08	7.846E-05	1.396E-07	6.729E-05	1.377E-07	1.776E-04	2.873E-07	1.191E-02	2.111E-04	1.830E-04	3.225E-07			3.529E-01	1.408E-01
M. musculus		2136	1.578E-03	34201	3.488E-03	65690	8.412E-03	1661771	6.457E-01	11749	3.143E-03	585	1.337E-03	801	1.691E-03	705	1.996E-03	1765	1.497E-03	122322	1.695E-02	1813	1.626E-03	3791399	3.529E-01		
10090	9503051	2.248E-04	3.546E-07	3.599E-03	1.255E-05	6.913E-03	5.815E-05	1.749E-01	1.129E-01	1.236E-03	3.885E-06	6.156E-05	8.231E-08	8.429E-05	1.425E-07	7.419E-05	1.481E-07	1.857E-04	2.780E-07	1.287E-02	2.182E-04	1.908E-04	3.101E-07	3.990E-01	1.408E-01		

The following table is the comparison of 13×13 pairs of model organisms by using the first filtering strategy. The technique of counting the number of 9-mers in the overlap (including duplicates) of both proteomes is used here. See next page for the complete table.

Proteome Name	Taxonomic ID	601		3702		6239		10116		36329		50339		85962		115711		224308		7227		243277		9606		10090	
Taxonomic ID	# of unique 9-mers	1369302		11127054		9360910		2827972		3948101		440589		479483		354710		1189168		10741502		1124273		15367806		12842290	
S. typhi				5642	5.071E-04	2566	2.741E-04	1391	4.919E-04	1062	2.690E-04	586	1.330E-03	2634	5.493E-03	1544	4.353E-03	5940	4.995E-03	5394	5.022E-04	52015	4.627E-02	4167	2.712E-04	3475	2.706E-04
601	1369302			4.120E-03	2.089E-06	1.874E-03	5.137E-07	1.016E-03	4.997E-07	7.756E-04	2.086E-07	4.280E-04	5.692E-07	1.924E-03	1.057E-05	1.128E-03	4.908E-06	4.338E-03	2.167E-05	3.939E-03	1.978E-06	3.799E-02	1.757E-03	3.043E-03	8.252E-07	2.538E-03	6.867E-07
A. thaliana		5642	4.120E-03			48163	5.145E-03	32190	1.138E-02	34557	8.753E-03	1183	2.685E-03	2363	4.928E-03	2205	6.216E-03	5430	4.566E-03	73240	6.818E-03	5194	4.620E-03	80252	5.222E-03	75459	5.876E-03
3702	11127054	5.071E-04	2.089E-06			4.328E-03	2.227E-05	2.893E-03	3.293E-05	3.106E-03	2.718E-05	1.063E-04	2.855E-07	2.124E-04	1.047E-06	1.982E-04	1.232E-06	4.880E-04	2.228E-06	6.582E-03	4.488E-05	4.668E-04	2.157E-06	7.212E-03	3.766E-05	6.782E-03	3.985E-05
C. elegans		2566	1.874E-03	48163	4.328E-03			61552	2.177E-02	23267	5.893E-03	657	1.491E-03	1003	2.092E-03	923	2.602E-03	2269	1.908E-03	128253	1.194E-02	2634	2.343E-03	143961	9.368E-03	133331	1.038E-02
6239	9360910	2.741E-04	5.137E-07	5.145E-03	2.227E-05			6.575E-03	1.431E-04	2.486E-03	1.465E-05	7.019E-05	1.047E-07	1.071E-04	2.241E-07	9.860E-05	2.566E-07	2.424E-04	4.625E-07	1.370E-02	1.636E-04	2.814E-04	6.592E-07	1.538E-02	1.441E-04	1.424E-02	1.479E-04
R. norvegicus		1391	1.016E-03	32190	2.893E-03	61552	6.575E-03			12030	3.047E-03	338	7.672E-04	500	1.043E-03	454	1.280E-03	1287	1.082E-03	109610	1.020E-02	1321	1.175E-03	1697757	1.105E-01	2170610	1.690E-01
10116	2827972	4.919E-04	4.997E-07	1.138E-02	3.293E-05	2.177E-02	1.431E-04			4.254E-03	1.296E-05	1.195E-04	9.169E-08	1.768E-04	1.844E-07	1.605E-04	2.055E-07	4.551E-04	4.925E-07	3.876E-02	3.955E-04	4.671E-04	5.489E-07	6.003E-01	6.632E-02	7.676E-01	1.297E-01
P. falciparum		1062	7.756E-04	34557	3.106E-03	23267	2.486E-03	12030	4.254E-03			387	8.784E-04	598	1.247E-03	504	1.421E-03	1010	8.493E-04	36929	3.438E-03	1064	9.464E-04	31568	2.054E-03	30823	2.400E-03
36329	3948101	2.690E-04	2.086E-07	8.753E-03	2.718E-05	5.893E-03	1.465E-05	3.047E-03	1.296E-05			9.802E-05	8.610E-08	1.515E-04	1.889E-07	1.277E-04	1.814E-07	2.558E-04	2.173E-07	9.354E-03	3.216E-05	2.695E-04	2.550E-07	7.996E-03	1.642E-05	7.807E-03	1.874E-05
T. volcanium		586	4.280E-04	1183	1.063E-04	657	7.019E-05	338	1.195E-04	387	9.802E-05			248	5.172E-04	144	4.060E-04	845	7.106E-04	880	8.193E-05	388	3.451E-04	936	6.091E-05	843	6.564E-05
50339	440589	1.330E-03	5.692E-07	2.685E-03	2.855E-07	1.491E-03	1.047E-07	7.672E-04	9.169E-08	8.784E-04	8.610E-08			5.629E-04	2.911E-07	3.268E-04	1.327E-07	1.918E-03	1.363E-06	1.997E-03	1.636E-07	8.806E-04	3.039E-07	2.124E-03	1.294E-07	1.913E-03	1.256E-07
H. pylori		2634	1.924E-03	2363	2.124E-04	1003	1.071E-04	500	1.768E-04	598	1.515E-04	248	5.629E-04			899	2.534E-03	2232	1.877E-03	1239	1.153E-04	2760	2.455E-03	1261	8.205E-05	1249	9.726E-05
85962	479483	5.493E-03	1.057E-05	4.928E-03	1.047E-06	2.092E-03	2.241E-07	1.043E-03	1.844E-07	1.247E-03	1.889E-07	5.172E-04	2.911E-07			1.875E-03	4.752E-06	4.655E-03	8.737E-06	2.584E-03	2.981E-07	5.756E-03	1.413E-05	2.630E-03	2.158E-07	2.605E-03	2.533E-07
C. pneumoniae		1544	1.128E-03	2205	1.982E-04	923	9.860E-05	454	1.605E-04	504	1.277E-04	144	3.268E-04	899	1.875E-03			1490	1.253E-03	1437	1.338E-04	1412	1.256E-03	1440	9.370E-05	1232	9.593E-05
115711	354710	4.353E-03	4.908E-06	6.216E-03	1.232E-06	2.602E-03	2.566E-07	1.280E-03	2.055E-07	1.421E-03	1.814E-07	4.060E-04	1.327E-07	2.534E-03	4.752E-06			4.201E-03	5.263E-06	4.051E-03	5.420E-07	3.981E-03	4.999E-06	4.060E-03	3.804E-07	3.473E-03	3.332E-07
B. subtilis		5940	4.338E-03	5430	4.880E-04	2269	2.424E-04	1287	4.551E-04	1010	2.558E-04	845	1.918E-03	2232	4.655E-03	1490	4.201E-03			3087	2.874E-04	5360	4.768E-03	3245	2.112E-04	2939	2.289E-04
224308	1189168	4.995E-03	2.167E-05	4.566E-03	2.228E-06	1.908E-03	4.625E-07	1.082E-03	4.925E-07	8.493E-04	2.173E-07	7.106E-04	1.363E-06	1.877E-03	8.737E-06	1.253E-03	5.263E-06			2.596E-03	7.460E-07	4.507E-03	2.149E-05	2.729E-03	5.762E-07	2.471E-03	5.656E-07
D. melanogaster		5394	3.939E-03	73240	6.582E-03	128253	1.370E-02	109610	3.876E-02	36929	9.354E-03	880	1.997E-03	1239	2.584E-03	1437	4.051E-03	3087	2.596E-03			4767	4.240E-03	280626	1.826E-02	257459	2.005E-02
7227	10741502	5.022E-04	1.978E-06	6.818E-03	4.488E-05	1.194E-02	1.636E-04	1.020E-02	3.955E-04	3.438E-03	3.216E-05	8.193E-05	1.636E-07	1.153E-04	2.981E-07	1.338E-04	5.420E-07	2.874E-04	7.460E-07			4.438E-04	1.882E-06	2.613E-02	4.771E-04	2.397E-02	4.805E-04
V. cholerae		52015	3.799E-02	5194	4.668E-04	2634	2.814E-04	1321	4.671E-04	1064	2.695E-04	388	8.806E-04	2760	5.756E-03	1412	3.981E-03	5360	4.507E-03	4767	4.438E-04			3591	2.337E-04	3334	2.596E-04
243277	1124273	4.627E-02	1.757E-03	4.620E-03	2.157E-06	2.343E-03	6.592E-07	1.175E-03	5.489E-07	9.464E-04	2.550E-07	3.451E-04	3.039E-07	2.455E-03	1.413E-05	1.256E-03	4.999E-06	4.768E-03	2.149E-05	4.240E-03	1.882E-06			3.194E-03	7.464E-07	2.965E-03	7.699E-07
H. sapiens		4167	3.043E-03	80252	7.212E-03	143961	1.538E-02	1697757	6.003E-01	31568	7.996E-03	936	2.124E-03	1261	2.630E-03	1440	4.060E-03	3245	2.729E-03	280626	2.613E-02	3591	3.194E-03			5664630	4.411E-01
9606	15367806	2.712E-04	8.252E-07	5.222E-03	3.766E-05	9.368E-03	1.441E-04	1.105E-01	6.632E-02	2.054E-03	1.642E-05	6.091E-05	1.294E-07	8.205E-05	2.158E-07	9.370E-05	3.804E-07	2.112E-04	5.762E-07	1.826E-02	4.771E-04	2.337E-04	7.464E-07			3.686E-01	1.626E-01
M. musculus		3475	2.538E-03	75459	6.782E-03	133331	1.424E-02	2170610	7.676E-01	30823	7.807E-03	843	1.913E-03	1249	2.605E-03	1232	3.473E-03	2939	2.471E-03	257459	2.397E-02	3334	2.965E-03	5664630	3.686E-01		
10090	12842290	2.706E-04	6.867E-07	5.876E-03	3.985E-05	1.038E-02	1.479E-04	1.690E-01	1.297E-01	2.400E-03	1.874E-05	6.564E-05	1.256E-07	9.726E-05	2.533E-07	9.593E-05	3.332E-07	2.289E-04	5.656E-07	2.005E-02	4.805E-04	2.596E-04	7.699E-07	4.411E-01	1.626E-01		

The following table is the comparison of 13×13 pairs of model organisms by using the second filtering strategy. The technique of counting the unique 9-mers in the overlap of both proteomes is used here. See next page for the complete table.

Proteome Name	Taxonomic ID	601		3702		6239		10116		36329		50339		85962		115711		224308		7227		243277		9606		10090	
Taxonomic ID	**# of unique 9-mers**	1353169		9804087		7806071		2542425		3735883		437176		469880		344862		1178837		7208445		1106020		10670465		9444529	
S. typhi				3833	3.910E-04	1711	2.192E-04	1038	4.083E-04	894	2.393E-04	536	1.226E-03	2495	5.310E-03	1443	4.184E-03	5406	4.586E-03	2233	3.098E-04	50372	4.554E-02	2602	2.439E-04	2124	2.249E-04
601	1353169			2.833E-03	1.107E-06	1.264E-03	2.772E-07	7.671E-04	3.132E-07	6.607E-04	1.581E-07	3.961E-04	4.856E-07	1.844E-03	9.790E-06	1.066E-03	4.462E-06	3.995E-03	1.832E-05	1.650E-03	5.112E-07	3.723E-02	1.695E-03	1.923E-03	4.689E-07	1.570E-03	3.530E-07
A. thaliana		3833	2.833E-03			23327	2.988E-03	14052	5.527E-03	12855	3.441E-03	778	1.780E-03	1532	3.260E-03	1370	3.973E-03	3422	2.903E-03	30143	4.182E-03	3334	3.014E-03	35867	3.361E-03	34134	3.614E-03
3702	9804087	3.910E-04	1.107E-06			2.379E-03	7.110E-06	1.433E-03	7.922E-06	1.311E-03	4.512E-06	7.935E-05	1.412E-07	1.563E-04	5.095E-07	1.397E-04	5.551E-07	3.490E-04	1.013E-06	3.075E-03	1.286E-05	3.401E-04	1.025E-06	3.658E-03	1.230E-05	3.482E-03	1.258E-05
C. elegans		1711	1.264E-03	23327	2.379E-03			34080	1.340E-02	9759	2.612E-03	530	1.212E-03	622	1.324E-03	532	1.543E-03	1418	1.203E-03	63265	8.777E-03	1647	1.489E-03	68022	6.375E-03	65538	6.939E-03
6239	7806071	2.192E-04	2.772E-07	2.988E-03	7.110E-06			4.366E-03	5.852E-05	1.250E-03	3.266E-06	6.790E-05	8.231E-08	7.968E-05	1.055E-07	6.815E-05	1.051E-07	1.817E-04	2.185E-07	8.105E-03	7.113E-05	2.110E-04	3.142E-07	8.714E-03	5.555E-05	8.396E-03	5.826E-05
R. norvegicus		1038	7.671E-04	14052	1.433E-03	34080	4.366E-03			5708	1.528E-03	269	6.153E-04	393	8.364E-04	329	9.540E-04	959	8.135E-04	54798	7.602E-03	871	7.875E-04	1207789	1.132E-01	1634994	1.731E-01
10116	2542425	4.083E-04	3.132E-07	5.527E-03	7.922E-06	1.340E-02	5.852E-05			2.245E-03	3.430E-06	1.058E-04	6.510E-08	1.546E-04	1.293E-07	1.294E-04	1.235E-07	3.772E-04	3.069E-07	2.155E-02	1.638E-04	3.426E-04	2.698E-07	4.751E-01	5.377E-02	6.431E-01	1.113E-01
P. falciparum		894	6.607E-04	12855	1.311E-03	9759	1.250E-03	5708	2.245E-03			344	7.869E-04	508	1.081E-03	417	1.209E-03	823	6.981E-04	11592	1.608E-03	844	7.631E-04	12003	1.125E-03	11677	1.236E-03
36329	3735883	2.393E-04	1.581E-07	3.441E-03	4.512E-06	2.612E-03	3.266E-06	1.528E-03	3.430E-06			9.208E-05	7.245E-08	1.360E-04	1.470E-07	1.116E-04	1.350E-07	2.203E-04	1.538E-07	3.103E-03	4.990E-06	2.259E-04	1.724E-07	3.213E-03	3.614E-06	3.126E-03	3.864E-06
T. volcanium		536	3.961E-04	778	7.935E-05	530	6.790E-05	269	1.058E-04	344	9.208E-05			234	4.980E-04	135	3.915E-04	751	6.371E-04	517	7.172E-05	337	3.047E-04	626	5.867E-05	582	6.162E-05
50339	437176	1.226E-03	4.856E-07	1.780E-03	1.412E-07	1.212E-03	8.231E-08	6.153E-04	6.510E-08	7.869E-04	7.245E-08			5.353E-04	2.666E-07	3.088E-04	1.209E-07	1.718E-03	1.094E-06	1.183E-03	8.482E-08	7.709E-04	2.349E-07	1.432E-03	8.401E-08	1.331E-03	8.204E-08
H. pylori		2495	1.844E-03	1532	1.563E-04	622	7.968E-05	393	1.546E-04	508	1.360E-04	234	5.353E-04			878	2.546E-03	2028	1.720E-03	722	1.002E-04	2234	2.020E-03	838	7.853E-05	800	8.471E-05
85962	469880	5.310E-03	9.790E-06	3.260E-03	5.095E-07	1.324E-03	1.055E-07	8.364E-04	1.293E-07	1.081E-03	1.470E-07	4.980E-04	2.666E-07			1.869E-03	4.757E-06	4.316E-03	7.425E-06	1.537E-03	1.539E-07	4.754E-03	9.603E-06	1.783E-03	1.401E-07	1.703E-03	1.442E-07
C. pneumoniae		1443	1.066E-03	1370	1.397E-04	532	6.815E-05	329	1.294E-04	417	1.116E-04	135	3.088E-04	878	1.869E-03			1374	1.166E-03	659	9.142E-05	1273	1.151E-03	717	6.719E-05	702	7.433E-05
115711	344862	4.184E-03	4.462E-06	3.973E-03	5.551E-07	1.543E-03	1.051E-07	9.540E-04	1.235E-07	1.209E-03	1.350E-07	3.915E-04	1.209E-07	2.546E-03	4.757E-06			3.984E-03	4.644E-06	1.911E-03	1.747E-07	3.691E-03	4.249E-06	2.079E-03	1.397E-07	2.036E-03	1.513E-07
B. subtilis		5406	3.995E-03	3422	3.490E-04	1418	1.817E-04	959	3.772E-04	823	2.203E-04	751	1.718E-03	2028	4.316E-03	1374	3.984E-03			1583	2.196E-04	4440	4.014E-03	1904	1.784E-04	1762	1.866E-04
224308	1178837	4.586E-03	1.832E-05	2.903E-03	1.013E-06	1.203E-03	2.185E-07	8.135E-04	3.069E-07	6.981E-04	1.538E-07	6.371E-04	1.094E-06	1.720E-03	7.425E-06	1.166E-03	4.644E-06			1.343E-03	2.949E-07	3.766E-03	1.512E-05	1.615E-03	2.882E-07	1.495E-03	2.789E-07
D. melanogaster		2233	1.650E-03	30143	3.075E-03	63265	8.105E-03	54798	2.155E-02	11592	3.103E-03	517	1.183E-03	722	1.537E-03	659	1.911E-03	1583	1.343E-03			1683	1.522E-03	127647	1.196E-02	121846	1.290E-02
7227	7208445	3.098E-04	5.112E-07	4.182E-03	1.286E-05	8.777E-03	7.113E-05	7.602E-03	1.638E-04	1.608E-03	4.990E-06	7.172E-05	8.482E-08	1.002E-04	1.539E-07	9.142E-05	1.747E-07	2.196E-04	2.949E-07			2.335E-04	3.553E-07	1.771E-02	2.118E-04	1.690E-02	2.181E-04
V. cholerae		50372	3.723E-02	3334	3.401E-04	1647	2.110E-04	871	3.426E-04	844	2.259E-04	337	7.709E-04	2234	4.754E-03	1273	3.691E-03	4440	3.766E-03	1683	2.335E-04			1957	1.834E-04	1809	1.915E-04
243277	1106020	4.554E-02	1.695E-03	3.014E-03	1.025E-06	1.489E-03	3.142E-07	7.875E-04	2.698E-07	7.631E-04	1.724E-07	3.047E-04	2.349E-07	2.020E-03	9.603E-06	1.151E-03	4.249E-06	4.014E-03	1.512E-05	1.522E-03	3.553E-07			1.769E-03	3.245E-07	1.636E-03	3.133E-07
H. sapiens		2602	1.923E-03	35867	3.658E-03	68022	8.714E-03	1207789	4.751E-01	12003	3.213E-03	626	1.432E-03	838	1.783E-03	717	2.079E-03	1904	1.615E-03	127637	1.771E-02	1957	1.769E-03			3759302	3.980E-01
9606	10670465	2.439E-04	4.689E-07	3.361E-03	1.230E-05	6.375E-03	5.555E-05	1.132E-01	5.377E-02	1.125E-03	3.614E-06	5.867E-05	8.401E-08	7.853E-05	1.401E-07	6.719E-05	1.397E-07	1.784E-04	2.882E-07	1.196E-02	2.118E-04	1.834E-04	3.245E-07			3.523E-01	1.402E-01
M. musculus		2124	1.570E-03	34134	3.482E-03	65538	8.396E-03	1634994	6.431E-01	11677	3.126E-03	582	1.331E-03	800	1.703E-03	702	2.036E-03	1762	1.495E-03	121846	1.690E-02	1809	1.636E-03	3759302	3.523E-01		
10090	9444529	2.249E-04	3.530E-07	3.614E-03	1.258E-05	6.939E-03	5.826E-05	1.731E-01	1.113E-01	1.236E-03	3.864E-06	6.162E-05	8.204E-08	8.471E-05	1.442E-07	7.433E-05	1.513E-07	1.866E-04	2.789E-07	1.290E-02	2.181E-04	1.915E-04	3.133E-07	3.980E-01	1.402E-01		

The following table is the comparison of 13×13 pairs of model organisms by using the second filtering strategy. The technique of counting the number of 9-mers in the overlap (including duplicates) of both proteomes is used here. See next page for the complete table.

Proteome Name	**Taxonomic ID**	601		3702		6239		10116		36329		50339		85962		115711		224308		7227		243277		9606		10090	
Taxonomic ID	**# of 9-mer occurrences**	1368473		11124178		9356423		2793097		3945141		440200		474271		346245		1188706		10709861		1114925		15224726		12721665	
S. typhi				5634	5.065E-04	2560	2.736E-04	1352	4.841E-04	1062	2.692E-04	586	1.331E-03	2633	5.552E-03	1533	4.428E-03	5937	4.995E-03	5332	4.979E-04	51911	4.656E-02	4149	2.725E-04	3437	2.702E-04
601	1368473			4.117E-03	2.085E-06	1.871E-03	5.118E-07	9.880E-04	4.782E-07	7.760E-04	2.089E-07	4.282E-04	5.700E-07	1.924E-03	1.068E-05	1.120E-03	4.960E-06	4.338E-03	2.167E-05	3.896E-03	1.940E-06	3.793E-02	1.766E-03	3.032E-03	8.262E-07	2.512E-03	6.785E-07
A. thaliana		5634	4.117E-03			48154	5.147E-03	31855	1.140E-02	34458	8.734E-03	1182	2.685E-03	2363	4.982E-03	2179	6.293E-03	5430	4.568E-03	73033	6.819E-03	5176	4.642E-03	80103	5.261E-03	75244	5.915E-03
3702	11124178	5.065E-04	2.085E-06			4.329E-03	2.228E-05	2.864E-03	3.266E-05	3.098E-03	2.706E-05	1.063E-04	2.853E-07	2.124E-04	1.058E-06	1.959E-04	1.233E-06	4.881E-04	2.230E-06	6.565E-03	4.477E-05	4.653E-04	2.160E-06	7.201E-03	3.789E-05	6.764E-03	4.001E-05
C. elegans		2560	1.871E-03	48154	4.329E-03			60956	2.182E-02	23214	5.884E-03	657	1.493E-03	1003	2.115E-03	922	2.663E-03	2269	1.909E-03	127955	1.195E-02	2624	2.354E-03	143581	9.431E-03	132852	1.044E-02
6239	9356423	2.736E-04	5.118E-07	5.147E-03	2.228E-05			6.515E-03	1.422E-04	2.481E-03	1.460E-05	7.022E-05	1.048E-07	1.072E-04	2.267E-07	9.854E-05	2.624E-07	2.425E-04	4.629E-07	1.368E-02	1.634E-04	2.804E-04	6.600E-07	1.535E-02	1.447E-04	1.420E-02	1.483E-04
R. norvegicus		1352	9.880E-04	31855	2.864E-03	60956	6.515E-03			11302	2.865E-03	332	7.542E-04	499	1.052E-03	436	1.259E-03	1255	1.056E-03	108335	1.012E-02	1267	1.136E-03	1672778	1.099E-01	2134784	1.678E-01
10116	2793097	4.841E-04	4.782E-07	1.140E-02	3.266E-05	2.182E-02	1.422E-04			4.046E-03	1.159E-05	1.189E-04	8.965E-08	1.787E-04	1.880E-07	1.561E-04	1.966E-07	4.493E-04	4.744E-07	3.879E-02	3.923E-04	4.536E-04	5.155E-07	5.989E-01	6.580E-02	7.643E-01	1.283E-01
P. falciparum		1062	7.760E-04	34458	3.098E-03	23214	2.481E-03	11302	4.046E-03			385	8.746E-04	596	1.257E-03	502	1.450E-03	1009	8.488E-04	36715	3.428E-03	1063	9.534E-04	31281	2.055E-03	30542	2.401E-03
36329	3945141	2.692E-04	2.089E-07	8.734E-03	2.706E-05	5.884E-03	1.460E-05	2.865E-03	1.159E-05			9.759E-05	8.535E-08	1.511E-04	1.898E-07	1.272E-04	1.845E-07	2.558E-04	2.171E-07	9.306E-03	3.190E-05	2.694E-04	2.569E-07	7.929E-03	1.629E-05	7.742E-03	1.859E-05
T. volcanium		586	4.282E-04	1182	1.063E-04	657	7.022E-05	332	1.189E-04	385	9.759E-05			248	5.229E-04	142	4.101E-04	842	7.083E-04	838	7.825E-05	388	3.480E-04	932	6.122E-05	834	6.556E-05
50339	440200	1.331E-03	5.700E-07	2.685E-03	2.853E-07	1.493E-03	1.048E-07	7.542E-04	8.965E-08	8.746E-04	8.535E-08			5.634E-04	2.946E-07	3.226E-04	1.323E-07	1.913E-03	1.355E-06	1.904E-03	1.490E-07	8.814E-04	3.067E-07	2.117E-03	1.296E-07	1.895E-03	1.242E-07
H. pylori		2633	1.924E-03	2363	2.124E-04	1003	1.072E-04	499	1.787E-04	596	1.511E-04	248	5.634E-04			898	2.594E-03	2227	1.873E-03	1239	1.157E-04	2754	2.470E-03	1254	8.237E-05	1240	9.747E-05
85962	474271	5.552E-03	1.068E-05	4.982E-03	1.058E-06	2.115E-03	2.267E-07	1.052E-03	1.880E-07	1.257E-03	1.898E-07	5.229E-04	2.946E-07			1.893E-03	4.911E-06	4.696E-03	8.797E-06	2.612E-03	3.022E-07	5.807E-03	1.434E-05	2.644E-03	2.178E-07	2.615E-03	2.548E-07
C. pneumoniae		1533	1.120E-03	2179	1.959E-04	922	9.854E-05	436	1.561E-04	502	1.272E-04	142	3.226E-04	898	1.893E-03			1473	1.239E-03	1434	1.339E-04	1392	1.249E-03	1431	9.399E-05	1228	9.653E-05
115711	346245	4.428E-03	4.960E-06	6.293E-03	1.233E-06	2.663E-03	2.624E-07	1.259E-03	1.966E-07	1.450E-03	1.845E-07	4.101E-04	1.323E-07	2.594E-03	4.911E-06			4.254E-03	5.272E-06	4.142E-03	5.545E-07	4.020E-03	5.019E-06	4.133E-03	3.885E-07	3.547E-03	3.423E-07
B. subtilis		5937	4.338E-03	5430	4.881E-04	2269	2.425E-04	1255	4.493E-04	1009	2.558E-04	842	1.913E-03	2227	4.696E-03	1473	4.254E-03			3040	2.839E-04	5337	4.787E-03	3237	2.126E-04	2929	2.302E-04
224308	1188706	4.995E-03	2.167E-05	4.568E-03	2.230E-06	1.909E-03	4.629E-07	1.056E-03	4.744E-07	8.488E-04	2.171E-07	7.083E-04	1.355E-06	1.873E-03	8.797E-06	1.239E-03	5.272E-06			2.557E-03	7.259E-07	4.490E-03	2.149E-05	2.723E-03	5.790E-07	2.464E-03	5.673E-07
D. melanogaster		5332	3.896E-03	73033	6.565E-03	127955	1.368E-02	108335	3.879E-02	36715	9.306E-03	838	1.904E-03	1239	2.612E-03	1434	4.142E-03	3040	2.557E-03			4752	4.262E-03	279291	1.834E-02	255896	2.011E-02
7227	10709861	4.979E-04	1.940E-06	6.819E-03	4.477E-05	1.195E-02	1.634E-04	1.012E-02	3.923E-04	3.428E-03	3.190E-05	7.825E-05	1.490E-07	1.157E-04	3.022E-07	1.339E-04	5.545E-07	2.839E-04	7.259E-07			4.437E-04	1.891E-06	2.608E-02	4.784E-04	2.389E-02	4.806E-04
V. cholerae		51911	3.793E-02	5176	4.653E-04	2624	2.804E-04	1267	4.536E-04	1063	2.694E-04	388	8.814E-04	2754	5.807E-03	1392	4.020E-03	5337	4.490E-03	4752	4.437E-04			3576	2.349E-04	3307	2.600E-04
243277	1114925	4.656E-02	1.766E-03	4.642E-03	2.160E-06	2.354E-03	6.600E-07	1.136E-03	5.155E-07	9.534E-04	2.569E-07	3.480E-04	3.067E-07	2.470E-03	1.434E-05	1.249E-03	5.019E-06	4.787E-03	2.149E-05	4.262E-03	1.891E-06			3.207E-03	7.534E-07	2.966E-03	7.710E-07
H. sapiens		4149	3.032E-03	80103	7.201E-03	143581	1.535E-02	1672778	5.989E-01	31281	7.929E-03	932	2.117E-03	1254	2.644E-03	1431	4.133E-03	3237	2.723E-03	279291	2.608E-02	3576	3.207E-03			5603713	4.405E-01
9606	15224726	2.725E-04	8.262E-07	5.261E-03	3.789E-05	9.431E-03	1.447E-04	1.099E-01	6.580E-02	2.055E-03	1.629E-05	6.122E-05	1.296E-07	8.237E-05	2.178E-07	9.399E-05	3.885E-07	2.126E-04	5.790E-07	1.834E-02	4.784E-04	2.349E-04	7.534E-07			3.681E-01	1.621E-01
M. musculus		3437	2.512E-03	75244	6.764E-03	132852	1.420E-02	2134784	7.643E-01	30542	7.742E-03	834	1.895E-03	1240	2.615E-03	1228	3.547E-03	2929	2.464E-03	255896	2.389E-02	3307	2.966E-03	5603713	3.681E-01		
10090	12721665	2.702E-04	6.785E-07	5.915E-03	4.001E-05	1.044E-02	1.483E-04	1.678E-01	1.283E-01	2.401E-03	1.859E-05	6.556E-05	1.242E-07	9.747E-05	2.548E-07	9.653E-05	3.423E-07	2.302E-04	5.673E-07	2.011E-02	4.806E-04	2.600E-04	7.710E-07	4.405E-01	1.621E-01		

The following table is the comparison of 13×13 pairs of model organisms using the third filtering strategy. The technique of counting the unique 9-mers in the overlap of both proteomes is used here. See next page for the complete table.

Proteome Name	**Taxonomic ID**	601		3702		6239		10116		36329		50339		85962		115711		224308		7227		243277		9606		10090	
Taxonomic ID	**# of unique 9-mers**	1353953		9805006		7808774		2572460		3738467		437503		473682		353071		1179260		7216967		1115183		10739022		9499769	
S. typhi				3839	3.915E-04	1717	2.199E-04	1073	4.171E-04	894	2.391E-04	536	1.225E-03	2496	5.269E-03	1452	4.112E-03	5409	4.587E-03	2235	3.097E-04	50474	4.526E-02	2613	2.433E-04	2136	2.248E-04
601	1353953			2.835E-03	1.110E-06	1.268E-03	2.788E-07	7.925E-04	3.306E-07	6.603E-04	1.579E-07	3.959E-04	4.850E-07	1.843E-03	9.714E-06	1.072E-03	4.410E-06	3.995E-03	1.832E-05	1.651E-03	5.112E-07	3.728E-02	1.687E-03	1.930E-03	4.696E-07	1.578E-03	3.547E-07
A. thaliana		3839	2.835E-03			23329	2.988E-03	14200	5.520E-03	12930	3.459E-03	779	1.781E-03	1532	3.234E-03	1383	3.917E-03	3422	2.902E-03	30158	4.179E-03	3347	3.001E-03	35912	3.344E-03	34201	3.600E-03
3702	9805006	3.915E-04	1.110E-06			2.379E-03	7.108E-06	1.448E-03	7.994E-06	1.319E-03	4.561E-06	7.945E-05	1.415E-07	1.562E-04	5.053E-07	1.411E-04	5.525E-07	3.490E-04	1.013E-06	3.076E-03	1.285E-05	3.414E-04	1.025E-06	3.663E-03	1.225E-05	3.488E-03	1.256E-05
C. elegans		1717	1.268E-03	23329	2.379E-03			34388	1.337E-02	9798	2.621E-03	530	1.211E-03	622	1.313E-03	533	1.510E-03	1418	1.202E-03	63286	8.769E-03	1653	1.482E-03	68134	6.345E-03	65690	6.915E-03
6239	7808774	2.199E-04	2.788E-07	2.988E-03	7.108E-06			4.404E-03	5.887E-05	1.255E-03	3.289E-06	6.787E-05	8.222E-08	7.965E-05	1.046E-07	6.826E-05	1.030E-07	1.816E-04	2.184E-07	8.104E-03	7.107E-05	2.117E-04	3.138E-07	8.725E-03	5.536E-05	8.412E-03	5.817E-05
R. norvegicus		1073	7.925E-04	14200	1.448E-03	34388	4.404E-03			5855	1.566E-03	274	6.263E-04	393	8.297E-04	347	9.828E-04	985	8.353E-04	55280	7.660E-03	922	8.268E-04	1224570	1.140E-01	1661768	1.749E-01
10116	2572460	4.171E-04	3.306E-07	5.520E-03	7.994E-06	1.337E-02	5.887E-05			2.276E-03	3.565E-06	1.065E-04	6.671E-08	1.528E-04	1.268E-07	1.349E-04	1.326E-07	3.829E-04	3.198E-07	2.149E-02	1.646E-04	3.584E-04	2.963E-07	4.760E-01	5.428E-02	6.460E-01	1.130E-01
P. falciparum		894	6.603E-04	12930	1.319E-03	9798	1.255E-03	5855	2.276E-03			344	7.863E-04	510	1.077E-03	419	1.187E-03	823	6.979E-04	11651	1.614E-03	845	7.577E-04	12069	1.124E-03	11747	1.237E-03
36329	3738467	2.391E-04	1.579E-07	3.459E-03	4.561E-06	2.621E-03	3.289E-06	1.566E-03	3.565E-06			9.202E-05	7.235E-08	1.364E-04	1.469E-07	1.121E-04	1.330E-07	2.201E-04	1.536E-07	3.117E-03	5.031E-06	2.260E-04	1.713E-07	3.228E-03	3.628E-06	3.142E-03	3.886E-06
T. volcanium		536	3.959E-04	779	7.945E-05	530	6.787E-05	274	1.065E-04	344	9.202E-05			234	4.940E-04	136	3.852E-04	754	6.394E-04	518	7.178E-05	337	3.022E-04	629	5.857E-05	585	6.158E-05
50339	437503	1.225E-03	4.850E-07	1.781E-03	1.415E-07	1.211E-03	8.222E-08	6.263E-04	6.671E-08	7.863E-04	7.235E-08			5.349E-04	2.642E-07	3.109E-04	1.197E-07	1.723E-03	1.102E-06	1.184E-03	8.498E-08	7.703E-04	2.328E-07	1.438E-03	8.421E-08	1.337E-03	8.234E-08
H. pylori		2496	1.843E-03	1532	1.562E-04	622	7.965E-05	393	1.528E-04	510	1.364E-04	234	5.349E-04			879	2.490E-03	2033	1.724E-03	722	1.000E-04	2239	2.008E-03	843	7.850E-05	801	8.432E-05
85962	473682	5.269E-03	9.714E-06	3.234E-03	5.053E-07	1.313E-03	1.046E-07	8.297E-04	1.268E-07	1.077E-03	1.469E-07	4.940E-04	2.642E-07			1.856E-03	4.620E-06	4.292E-03	7.399E-06	1.524E-03	1.525E-07	4.727E-03	9.490E-06	1.780E-03	1.397E-07	1.691E-03	1.426E-07
C. pneumoniae		1452	1.072E-03	1383	1.411E-04	533	6.826E-05	347	1.349E-04	419	1.121E-04	136	3.109E-04	879	1.856E-03			1390	1.179E-03	661	9.159E-05	1289	1.156E-03	723	6.732E-05	705	7.421E-05
115711	353071	4.112E-03	4.410E-06	3.917E-03	5.525E-07	1.510E-03	1.030E-07	9.828E-04	1.326E-07	1.187E-03	1.330E-07	3.852E-04	1.197E-07	2.490E-03	4.620E-06			3.937E-03	4.640E-06	1.872E-03	1.715E-07	3.651E-03	4.220E-06	2.048E-03	1.379E-07	1.997E-03	1.482E-07
B. subtilis		5409	3.995E-03	3422	3.490E-04	1418	1.816E-04	985	3.829E-04	823	2.201E-04	754	1.723E-03	2033	4.292E-03	1390	3.937E-03			1583	2.193E-04	4460	3.999E-03	1908	1.777E-04	1765	1.858E-04
224308	1179260	4.587E-03	1.832E-05	2.902E-03	1.013E-06	1.202E-03	2.184E-07	8.353E-04	3.198E-07	6.979E-04	1.536E-07	6.394E-04	1.102E-06	1.724E-03	7.399E-06	1.179E-03	4.640E-06			1.342E-03	2.944E-07	3.782E-03	1.513E-05	1.618E-03	2.875E-07	1.497E-03	2.781E-07
D. melanogaster		2235	1.651E-03	30158	3.076E-03	63286	8.104E-03	55280	2.149E-02	11651	3.117E-03	518	1.184E-03	722	1.524E-03	661	1.872E-03	1583	1.342E-03			1686	1.512E-03	127941	1.191E-02	122322	1.288E-02
7227	7216967	3.097E-04	5.112E-07	4.179E-03	1.285E-05	8.769E-03	7.107E-05	7.660E-03	1.646E-04	1.614E-03	5.031E-06	7.178E-05	8.498E-08	1.000E-04	1.525E-07	9.159E-05	1.715E-07	2.193E-04	2.944E-07			2.336E-04	3.532E-07	1.773E-02	2.112E-04	1.695E-02	2.182E-04
V. cholerae		50474	3.728E-02	3347	3.414E-04	1653	2.117E-04	922	3.584E-04	845	2.260E-04	337	7.703E-04	2239	4.727E-03	1289	3.651E-03	4460	3.782E-03	1686	2.336E-04			1966	1.831E-04	1813	1.908E-04
243277	1115183	4.526E-02	1.687E-03	3.001E-03	1.025E-06	1.482E-03	3.138E-07	8.268E-04	2.963E-07	7.577E-04	1.713E-07	3.022E-04	2.328E-07	2.008E-03	9.490E-06	1.156E-03	4.220E-06	3.999E-03	1.513E-05	1.512E-03	3.532E-07			1.763E-03	3.227E-07	1.626E-03	3.103E-07
H. sapiens		2613	1.930E-03	35912	3.663E-03	68134	8.725E-03	1224570	4.760E-01	12069	3.228E-03	629	1.438E-03	843	1.780E-03	723	2.048E-03	1908	1.618E-03	127941	1.773E-02	1966	1.763E-03			3791379	3.991E-01
9606	10739022	2.433E-04	4.696E-07	3.344E-03	1.225E-05	6.345E-03	5.536E-05	1.140E-01	5.428E-02	1.124E-03	3.628E-06	5.857E-05	8.421E-08	7.850E-05	1.397E-07	6.732E-05	1.379E-07	1.777E-04	2.875E-07	1.191E-02	2.112E-04	1.831E-04	3.227E-07			3.530E-01	1.409E-01
M. musculus		2136	1.578E-03	34201	3.488E-03	65690	8.412E-03	1661768	6.460E-01	11747	3.142E-03	585	1.337E-03	801	1.691E-03	705	1.997E-03	1765	1.497E-03	122322	1.695E-02	1813	1.626E-03	3791379	3.530E-01		
10090	9499769	2.248E-04	3.547E-07	3.600E-03	1.256E-05	6.915E-03	5.817E-05	1.749E-01	1.130E-01	1.237E-03	3.886E-06	6.158E-05	8.234E-08	8.432E-05	1.426E-07	7.421E-05	1.482E-07	1.858E-04	2.781E-07	1.288E-02	2.182E-04	1.908E-04	3.103E-07	3.991E-01	1.409E-01		

The following table is the comparison of 13×13 pairs of model organisms using the third filtering strategy. The technique of counting the number of 9-mers in the overlap (including duplicates) of both proteomes is used here. See next page for the complete table.

Proteome Name	**Taxonomic ID**	601		3702		6239		10116		36329		50339		85962		115711		224308		7227		243277		9606		10090	
Taxonomic ID	**# of 9-mer occurrences**	1369284		11127000		9360801		2826217		3947981		440580		479362		354548		1189150		10740972		1124125		15359070		12838611	
S. typhi				5642	5.071E-04	2566	2.741E-04	1391	4.922E-04	1062	2.690E-04	586	1.330E-03	2634	5.495E-03	1544	4.355E-03	5940	4.995E-03	5394	5.022E-04	52015	4.627E-02	4167	2.713E-04	3475	2.707E-04
601	1369284			4.120E-03	2.089E-06	1.874E-03	5.137E-07	1.016E-03	5.000E-07	7.756E-04	2.086E-07	4.280E-04	5.692E-07	1.924E-03	1.057E-05	1.128E-03	4.911E-06	4.338E-03	2.167E-05	3.939E-03	1.978E-06	3.799E-02	1.758E-03	3.043E-03	8.256E-07	2.538E-03	6.869E-07
A. thaliana		5642	4.120E-03			48163	5.145E-03	32190	1.139E-02	34557	8.753E-03	1183	2.685E-03	2363	4.929E-03	2205	6.219E-03	5430	4.566E-03	73240	6.819E-03	5194	4.620E-03	80252	5.225E-03	75459	5.878E-03
3702	11127000	5.071E-04	2.089E-06			4.328E-03	2.227E-05	2.893E-03	3.295E-05	3.106E-03	2.718E-05	1.063E-04	2.855E-07	2.124E-04	1.047E-06	1.982E-04	1.232E-06	4.880E-04	2.228E-06	6.582E-03	4.488E-05	4.668E-04	2.157E-06	7.212E-03	3.769E-05	6.782E-03	3.986E-05
C. elegans		2566	1.874E-03	48163	4.328E-03			61552	2.178E-02	23267	5.893E-03	657	1.491E-03	1003	2.092E-03	923	2.603E-03	2269	1.908E-03	128253	1.194E-02	2634	2.343E-03	143961	9.373E-03	133331	1.039E-02
6239	9360801	2.741E-04	5.137E-07	5.145E-03	2.227E-05			6.576E-03	1.432E-04	2.486E-03	1.465E-05	7.019E-05	1.047E-07	1.071E-04	2.242E-07	9.860E-05	2.567E-07	2.424E-04	4.625E-07	1.370E-02	1.636E-04	2.814E-04	6.593E-07	1.538E-02	1.441E-04	1.424E-02	1.479E-04
R. norvegicus		1391	1.016E-03	32190	2.893E-03	61552	6.576E-03			11611	2.941E-03	338	7.672E-04	500	1.043E-03	454	1.281E-03	1287	1.082E-03	109610	1.020E-02	1321	1.175E-03	1697242	1.105E-01	2170120	1.690E-01
10116	2826217	4.922E-04	5.000E-07	1.139E-02	3.295E-05	2.178E-02	1.432E-04			4.108E-03	1.208E-05	1.196E-04	9.175E-08	1.769E-04	1.845E-07	1.606E-04	2.057E-07	4.554E-04	4.929E-07	3.878E-02	3.958E-04	4.674E-04	5.493E-07	6.005E-01	6.636E-02	7.679E-01	1.298E-01
P. falciparum		1062	7.756E-04	34557	3.106E-03	23267	2.486E-03	11611	4.108E-03			387	8.784E-04	598	1.247E-03	504	1.422E-03	1010	8.493E-04	36929	3.438E-03	1064	9.465E-04	31422	2.046E-03	30706	2.392E-03
36329	3947981	2.690E-04	2.086E-07	8.753E-03	2.718E-05	5.893E-03	1.465E-05	2.941E-03	1.208E-05			9.802E-05	8.610E-08	1.515E-04	1.890E-07	1.277E-04	1.815E-07	2.558E-04	2.173E-07	9.354E-03	3.216E-05	2.695E-04	2.551E-07	7.959E-03	1.628E-05	7.778E-03	1.860E-05
T. volcanium		586	4.280E-04	1183	1.063E-04	657	7.019E-05	338	1.196E-04	387	9.802E-05			248	5.174E-04	144	4.062E-04	845	7.106E-04	880	8.193E-05	388	3.452E-04	936	6.094E-05	843	6.566E-05
50339	440580	1.330E-03	5.692E-07	2.685E-03	2.855E-07	1.491E-03	1.047E-07	7.672E-04	9.175E-08	8.784E-04	8.610E-08			5.629E-04	2.912E-07	3.268E-04	1.327E-07	1.918E-03	1.363E-06	1.997E-03	1.636E-07	8.807E-04	3.040E-07	2.124E-03	1.295E-07	1.913E-03	1.256E-07
H. pylori		2634	1.924E-03	2363	2.124E-04	1003	1.071E-04	500	1.769E-04	598	1.515E-04	248	5.629E-04			899	2.536E-03	2232	1.877E-03	1239	1.154E-04	2760	2.455E-03	1261	8.210E-05	1249	9.728E-05
85962	479362	5.495E-03	1.057E-05	4.929E-03	1.047E-06	2.092E-03	2.242E-07	1.043E-03	1.845E-07	1.247E-03	1.890E-07	5.174E-04	2.912E-07			1.875E-03	4.755E-06	4.656E-03	8.740E-06	2.585E-03	2.982E-07	5.758E-03	1.414E-05	2.631E-03	2.160E-07	2.606E-03	2.535E-07
C. pneumoniae		1544	1.128E-03	2205	1.982E-04	923	9.860E-05	454	1.606E-04	504	1.277E-04	144	3.268E-04	899	1.875E-03			1490	1.253E-03	1437	1.338E-04	1412	1.256E-03	1440	9.376E-05	1232	9.596E-05
115711	354548	4.355E-03	4.911E-06	6.219E-03	1.232E-06	2.603E-03	2.567E-07	1.281E-03	2.057E-07	1.422E-03	1.815E-07	4.062E-04	1.327E-07	2.536E-03	4.755E-06			4.203E-03	5.266E-06	4.053E-03	5.422E-07	3.983E-03	5.002E-06	4.062E-03	3.808E-07	3.475E-03	3.334E-07
B. subtilis		5940	4.338E-03	5430	4.880E-04	2269	2.424E-04	1287	4.554E-04	1010	2.558E-04	845	1.918E-03	2232	4.656E-03	1490	4.203E-03			3087	2.874E-04	5360	4.768E-03	3245	2.113E-04	2939	2.289E-04
224308	1189150	4.995E-03	2.167E-05	4.566E-03	2.228E-06	1.908E-03	4.625E-07	1.082E-03	4.929E-07	8.493E-04	2.173E-07	7.106E-04	1.363E-06	1.877E-03	8.740E-06	1.253E-03	5.266E-06			2.596E-03	7.461E-07	4.507E-03	2.149E-05	2.729E-03	5.765E-07	2.472E-03	5.658E-07
D. melanogaster		5394	3.939E-03	73240	6.582E-03	128253	1.370E-02	109610	3.878E-02	36929	9.354E-03	880	1.997E-03	1239	2.585E-03	1437	4.053E-03	3087	2.596E-03			4767	4.241E-03	280626	1.827E-02	257478	2.005E-02
7227	10740972	5.022E-04	1.978E-06	6.819E-03	4.488E-05	1.194E-02	1.636E-04	1.020E-02	3.958E-04	3.438E-03	3.216E-05	8.193E-05	1.636E-07	1.154E-04	2.982E-07	1.338E-04	5.422E-07	2.874E-04	7.461E-07			4.438E-04	1.882E-06	2.613E-02	4.774E-04	2.397E-02	4.807E-04
V. cholerae		52015	3.799E-02	5194	4.668E-04	2634	2.814E-04	1321	4.674E-04	1064	2.695E-04	388	8.807E-04	2760	5.758E-03	1412	3.983E-03	5360	4.507E-03	4767	4.438E-04			3591	2.338E-04	3334	2.597E-04
243277	1124125	4.627E-02	1.758E-03	4.620E-03	2.157E-06	2.343E-03	6.593E-07	1.175E-03	5.493E-07	9.465E-04	2.551E-07	3.452E-04	3.040E-07	2.455E-03	1.414E-05	1.256E-03	5.002E-06	4.768E-03	2.149E-05	4.241E-03	1.882E-06			3.194E-03	7.469E-07	2.966E-03	7.702E-07
H. sapiens		4167	3.043E-03	80252	7.212E-03	143961	1.538E-02	1697242	6.005E-01	31422	7.959E-03	936	2.124E-03	1261	2.631E-03	1440	4.062E-03	3245	2.729E-03	280626	2.613E-02	3591	3.194E-03			5664397	4.412E-01
9606	15359070	2.713E-04	8.256E-07	5.225E-03	3.769E-05	9.373E-03	1.441E-04	1.105E-01	6.636E-02	2.046E-03	1.628E-05	6.094E-05	1.295E-07	8.210E-05	2.160E-07	9.376E-05	3.808E-07	2.113E-04	5.765E-07	1.827E-02	4.774E-04	2.338E-04	7.469E-07			3.688E-01	1.627E-01
M. musculus		3475	2.538E-03	75459	6.782E-03	133331	1.424E-02	2170120	7.679E-01	30706	7.778E-03	843	1.913E-03	1249	2.606E-03	1232	3.475E-03	2939	2.472E-03	257478	2.397E-02	3334	2.966E-03	5664397	3.688E-01		
10090	12838611	2.707E-04	6.869E-07	5.878E-03	3.986E-05	1.039E-02	1.479E-04	1.690E-01	1.298E-01	2.392E-03	1.860E-05	6.566E-05	1.256E-07	9.728E-05	2.535E-07	9.596E-05	3.334E-07	2.289E-04	5.658E-07	2.005E-02	4.807E-04	2.597E-04	7.702E-07	4.412E-01	1.627E-01		

APPENDIX D

TIME FOR DATA PROCESSING

This appendix contains 3 tables giving CPU time information for the three pairs of organisms at the 9-mer level, one for each filtering strategy. Time information includes user CPU time, system CPU time and total time for either generating the peptide universes or counting the overlaps.

Table D.1: Time information for the 3 pairs of organisms at the 9-mer level using the first filtering strategy.

Proteome(s)	User CPU Time	System CPU Time	Total Time
85962	4.666	0.193	4.859
115711	3.394	0.156	3.550
7227	175.321	4.535	179.856
243277	11.573	0.416	11.989
9606	244.163	6.699	250.862
10090	204.120	4.761	208.881
85962,115711	2.800	0.134	2.934
115711,85962	3.160	0.121	3.281
7227,243277	27.340	1.442	28.782
243277,7227	70.300	1.336	71.636
9606,10090	169.922	5.495	175.417
10090,9606	194.482	6.279	200.761

Table D.2: Time information for the 3 pairs of organisms at the 9-mer level using the second filtering strategy.

Proteome(s)	User CPU Time	System CPU Time	Total Time
85962	4.827	0.282	5.109
115711	3.456	0.215	3.671
7227	177.048	5.722	182.77
243277	11.972	0.606	12.578
9606	245.941	8.821	254.762
			Continued...

Table D.2 – continued from previous page			
Proteome(s)	**User CPU Time**	**System CPU Time**	**Total Time**
10090	201.827	7.046	208.873
85962,115711	2.932	0.182	3.114
115711,85962	3.244	0.117	3.361
7227,243277	27.35	1.796	29.146
243277,7227	70.178	1.74	71.918
9606,10090	169.946	7.161	177.107
10090,9606	192.192	5.583	197.775

Table D.3: Time information for the 3 pairs of organisms at the 9-mer level using the third filtering strategy.

Proteome(s)	**User CPU Time**	**System CPU Time**	**Total Time**
85962	7.543	0.262	7.805
115711	5.506	0.224	5.730
7227	238.405	6.318	244.723
243277	18.080	0.584	18.664
9606	341.490	11.967	353.457
10090	278.348	7.864	286.212
85962,115711	3.261	0.192	3.453
115711,85962	3.463	0.152	3.615
7227,243277	33.517	2.098	35.615
243277,7227	71.023	1.510	72.533
9606,10090	195.456	5.800	201.256
10090,9606	220.496	6.823	227.319

APPENDIX E

VIRAL PROTEOME DESCRIPTIONS

11926.FASTA polyprotein from Human T-cell leukemia virus type I strain ATK (HTLV-I) (Accession number P03362, Name POL_HTL1A) downloaded from http://www.expasy.org/uniprot/ on November 9, 2005

11177.FASTA 4 proteins from Newcastle disease virus strain Australia-Victoria/32 (NDVA) downloaded from http://ca.expasy.org/uniprot/ on September 23, 2005.

12132.FASTA polyprotein from Human rhinovirus 89 (HRV-89) (Accession number P07210, Name POLG_HRV89) downloaded from http://www.ncbi.nlm.nih.gov/entrez on November 3, 2005

31915.FASTA polyprotein from Human enterovirus 70 (EV70) strain J670/71 (Accession number P32537, Name POLG_HE701) downloaded from http://www.ebi.uniprot.org/ on October 28, 2005

10581.FASTA 8 proteins from Human papillomavirus type 16 (HPV16)(Accession number K02718) downloaded from http://www.ebi.ac.uk/genomes/virus.html on May 5, 2005.

11105.FASTA polyprotein from Hepatitis C virus genotype 1b (Accession number P26663) (isolate BK) downloaded from http://www.ebi.uniprot.org/uniprot on December 21, 2005

11053.FASTA polyprotein from Dengue virus type 1 (Accession number U88536) downloaded from http://www.ebi.ac.uk/genomes/virus.html on November 14, 2004

11290.FASTA 6 proteins from Infectious hematopoietic necrosis virus(Accession number X89213) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

11089.FASTA polyprotein from Yellow fever virus (Accession number X03700) downloaded from http://www.ebi.ac.uk/genomes/virus.html on November 14, 2004 Sequences is identical to that for P03314 (YEFV1)

11079.FASTA polyprotein from Murray Valley encephalitis virus (Accession number AF161266) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

11075.FASTA polyprotein from Japanese encephalitis virus (strain Jaoars982) (Accession number P32886) downloaded from http://ca.expasy.org/uniprot on June 6, 2006

11082.FASTA polyprotein from West Nile virus (Accession number P06935) downloaded from http://ca.expasy.org/uniprot on June 6, 2006

11029.FASTA 2 proteins from Ross River virus (Accession number M20162) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

59301.FASTA 3 proteins from Mayaro virus (Accession number AF237947) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

11027.FASTA 2 proteins from O'nyong-nyong virus (Accession number AF079456) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

11593.FASTA RNA-dependent RNA polymerase from Crimean-Congo hemorrhagic fever virus (Accession number Q52NX4) downloaded from http://ca.expasy.org/uniprot on June 6, 2006

162145.FASTA 9 proteins from Human metapneumovirus (Accession number AF371337) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

12814.FASTA 10 proteins from Respiratory syncytial virus (Accession number U39661) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

93838.FASTA 10 proteins from Influenza A virus (A/Goose/Guangdong/1/96(H5N1)) (Accession numbers AF144300, AF144301, AF144302, AF144303, AF144304, AF144305, AF144306, AF144307) downloaded from http://www.ebi.ac.uk/genomes/virus.html on September 14, 2006

11216.FASTA 6 proteins from Human parainfluenza virus 3 (Accession number AB012132) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

11269.FASTA 7 proteins from Lake Victoria marburgvirus (Accession number Z12132) downloaded from http://www.ebi.ac.uk/genomes/virus.html on August 2, 2006

11137.FASTA polyprotein from Human coronavirus (strain 229E)(Accession number Q05002) (HCoV-229E) downloaded from http://ca.expasy.org/uniprot/Q05002 on June 15, 2006

227859.FASTA polyprotein from Human coronavirus strain SARS (HCoV-SARS) (SARS-CoV) (Accession number P59641, Name R1AB_CVHSA) downloaded from http://www.expasy.org/uniprot on November 16, 2005

APPENDIX F

SUMMARY OF ACTUAL AND EXPECTED OVERLAPS BETWEEN HUMAN AND HUMAN VIRUSES

This appendix contains 5 tables giving summary information for degrees of overlapping between the human proteome and human viruses. For each viral proteome and for k from 5 to 9, the following data are recorded: unique k-mers in viral proteome, viral occurrences (including duplicates) in human proteins, human proteins involved in overlap, number of overlapping k-mers (including duplicates) assuming random model, number of unique overlapping k-mers and occurrences of k-mers in viral proteome (including duplicates).

Table F.1: Table of data which is used to generate Figure 5.3. col 1: Proteome ID; col 2: unique 5-mers in viral proteome; col 3: viral occurrences (including duplicates) in human proteins; col 4: human proteins involved in overlap; col 5: number of overlapping 5-mers (including duplicates) assuming random model; col 6: number of unique overlapping 5-mers; col 7: occurrences of 5-mers in viral proteome (including duplicates).

Proteome	col 2	col 3	col 4	col 5	col 6	col 7
11926	892	15244	10238	4529	822	892
11177	1864	25691	14722	9473	1764	1866
12132	2156	24818	14457	10965	1961	2160
31915	2177	23484	13983	11061	1985	2179
12080	2203	24074	14115	11193	2016	2205
10581	2419	29802	15931	12284	2245	2420
11105	3000	48165	21094	15252	2753	3005
11053	3386	38449	19041	17195	3058	3388
11290	3393	51909	21178	17241	3170	3397
11089	3400	44621	20228	17291	3069	3407
11079	3422	43949	20434	17408	3091	3430
11075	3423	42147	20228	17398	3112	3428
11082	3424	43196	20291	17388	3096	3426
11029	3622	43621	20150	18433	3297	3632
59301	3663	43205	20116	18625	3349	3670
11027	3743	42712	20153	19016	3395	3747

Continued...

Table F.1 – continued from previous page						
Proteome	col 2	col 3	col 4	col 5	col 6	col 7
11593	3930	56902	22136	20000	3682	3941
162145	4120	54555	21531	20943	3779	4127
12814	4389	46317	20422	22358	4031	4406
93838	4412	46966	20909	22464	4036	4427
11216	4807	54534	21980	24447	4418	4818
11269	4808	68972	23857	24447	4439	4818
227859	7042	79925	25748	35856	6389	7069

Table F.2: Table of data which is used to generate Figure 5.4. col 1: Proteome ID; col 2: unique 6-mers in viral proteome; col 3: viral occurrences (including duplicates) in human proteins; col 4: human proteins involved in overlap; col 5: number of overlapping 6-mers (including duplicates) assuming random model; col 6: number of unique overlapping 6-mers; col 7: occurrences of 6-mers in viral proteome (including duplicates).

Proteome	col 2	col 3	col 4	col 5	col 6	col 7
11926	891	1092	928	226	320	891
11177	1862	1596	1436	472	622	1862
12132	2158	1545	1372	547	623	2159
31915	2176	1385	1259	551	580	2176
12080	2204	1523	1320	558	619	2204
10581	2412	2069	1757	610	739	2412
11105	3004	4383	3233	761	1037	3004
11053	3387	2329	2045	858	960	3387
11290	3391	4084	3141	859	1126	3391
11089	3406	2777	2406	863	1005	3406
11079	3429	2858	2491	869	1017	3429
11075	3427	2532	2223	868	976	3427
11082	3425	2701	2339	867	990	3425
11029	3613	2779	2367	915	1069	3613
59301	3666	2637	2292	929	1061	3668
11027	3744	2607	2224	948	1099	3744
11593	3940	3559	2960	998	1263	3940
						Continued...

Table F.2 – continued from previous page						
Proteome	col 2	col 3	col 4	col 5	col 6	col 7
162145	4117	3351	2805	1043	1245	4118
12814	4393	2765	2359	1113	1191	4396
93838	4408	2754	2404	1119	1191	4417
11216	4812	3288	2802	1218	1350	4812
11269	4811	5828	3809	1219	1497	4811
11137	6750	4047	3394	1710	1695	6753
227859	7064	7655	4523	1790	1952	7068

Table F.3: Table of data which is used to generate Figure 5.5. col 1: Proteome ID; col 2: unique 7-mers in viral proteome; col 3: viral occurrences (including duplicates) in human proteins; col 4: human proteins involved in overlap; col 5: number of overlapping 7-mers (including duplicates) assuming random model; col 6: number of unique overlapping 7-mers; col 7: occurrences of 7-mers in viral proteome (including duplicates).

Proteome	col 2	col 3	col 4	col 5	col 6	col 7
11926	890	120	90	11	57	890
11177	1858	80	68	23	51	1858
12132	2158	85	79	27	58	2158
31915	2173	77	73	27	49	2173
12080	2203	145	128	28	69	2203
10581	2404	167	125	30	102	2404
11105	3003	429	382	38	133	3003
11053	3386	163	157	43	98	3386
11290	3385	523	371	43	124	3385
11089	3405	172	161	43	96	3405
11079	3428	158	148	43	85	3428
11075	3426	151	147	43	91	3426
11082	3424	181	174	43	98	3424
11029	3595	188	173	45	119	3595
59301	3666	177	160	46	104	3666
11027	3741	190	179	47	114	3741
11593	3939	259	227	50	127	3939

Continued...

Table F.3 – continued from previous page						
Proteome	col 2	col 3	col 4	col 5	col 6	col 7
162145	4109	236	214	52	123	4109
12814	4386	189	162	55	107	4386
93838	4400	149	141	56	96	4407
11216	4806	202	179	61	126	4806
11269	4804	301	268	61	151	4804
11137	6751	251	235	85	156	6752
227859	7067	630	563	89	202	7067

Table F.4: Table of data which is used to generate Figure 5.6. col 1: Proteome ID; col 2: unique 8-mers in viral proteome; col 3: viral occurrences (including duplicates) in human proteins; col 4: human proteins involved in overlap; col 5: number of overlapping 8-mers (including duplicates) assuming random model; col 6: number of unique overlapping 8-mers; col 7: occurrences of 8-mers in viral proteome (including duplicates).

Proteome	col 2	col 3	col 4	col 5	col 6	col 7
11926	889	23	16	0.560	12	889
11177	1854	7	7	1.168	4	1854
12132	2157	6	6	1.359	5	2157
31915	2170	4	4	1.368	3	2170
12080	2202	17	16	1.388	11	2202
10581	2396	28	4	1.510	27	2396
11105	3002	29	28	1.892	11	3002
11053	3385	5	5	2.133	5	3385
11290	3379	125	100	2.129	14	3379
11089	3404	9	9	2.145	7	3404
11079	3427	10	10	2.160	7	3427
11075	3425	4	4	2.158	3	3425
11082	3423	7	7	2.157	6	3423
11029	3577	10	9	2.254	7	3577
59301	3664	16	16	2.309	8	3664
11027	3738	11	9	2.356	6	3738
11593	3938	25	21	2.482	15	3938
						Continued...

Table F.4 – continued from previous page						
Proteome	col 2	col 3	col 4	col 5	col 6	col 7
162145	4100	14	13	2.584	11	4100
12814	4376	10	9	2.758	8	4376
93838	4392	5	3	2.771	5	4397
11216	4800	21	17	3.025	13	4800
11269	4797	31	27	3.023	17	4797
11137	6751	15	13	4.255	9	6751
227859	7066	37	35	4.453	20	7066

Table F.5: Table of data which is used to generate Figure 5.7. col 1: Proteome ID; col 2: unique 9-mers in viral proteome; col 3: viral occurrences (including duplicates) in human proteins; col 4: human proteins involved in overlap; col 5: number of overlapping 9-mers (including duplicates) assuming random model; col 6: number of unique overlapping 9-mers; col 7: occurrences of 9-mers in viral proteome (including duplicates).

Proteome	col 2	col 3	col 4	col 5	col 6	col 7
11926	888	7	3	0.028	5	888
11177	1850	0	0	0.058	0	1850
12132	2156	0	0	0.067	0	2156
31915	2167	0	0	0.068	0	2167
12080	2201	1	1	0.069	1	2201
10581	2388	16	1	0.075	16	2388
11105	3001	0	0	0.094	0	3001
11053	3384	0	0	0.106	0	3384
11290	3373	13	11	0.106	4	3373
11089	3403	0	0	0.107	0	3403
11079	3426	0	0	0.108	0	3426
11075	3424	0	0	0.108	0	3424
11082	3422	0	0	0.108	0	3422
11029	3559	1	1	0.112	1	3559
59301	3662	0	0	0.115	0	3662
11027	3735	2	2	0.117	1	3735
11593	3937	4	4	0.124	2	3937
						Continued...

Table F.5 – continued from previous page						
Proteome	col 2	col 3	col 4	col 5	col 6	col 7
162145	4091	1	1	0.129	1	4091
12814	4366	1	1	0.137	1	4366
93838	4384	2	1	0.138	2	4387
11216	4794	4	1	0.151	1	4790
11137	6750	2	2	0.212	1	6750
227859	7065	2	1	0.222	2	7065

APPENDIX G

SUMMARY OF CHI-SQUARE ANALYSIS

This appendix contains 9 tables giving summary information for counts of overlaps/nonoverlaps between viral proteomes (HIV-1, HIV-2 or Influenza A virus) and the human proteome at different peptide lengths (i.e., 5-, 6- and 7-mer).

The HIV-1 proteome is broken up into 36 segments of 100 amino acids in length at the 5-mer level.

Table G.1: Table of counts of nonoverlapping 5-mers by dividing the HIV-1 proteome into evenly-sized (i.e., 100 amino acids) segments. N is the observed number of 5-mers in the segment with no overlap in the human proteome.

Position	N	**HIV-1 protein**
0001-0100	2	Gag poly (p17)
0101-0200	8	Gag poly (p17, p24)
0201-0300	10	Gag poly (p24)
0301-0400	18	Gag poly (p24, p2, p7, p1, p6)
0401-0500	9	Gag poly (p2, p7, p1, p6)
0501-0600	3	Gag-Pol (p17)
0601-0700	7	Gag-Pol (p17, p24)
0701-0800	11	Gag-Pol (p24)
0801-0900	17	Gag-Pol (p24, p2, p7, p1, p6)
0901-1000	13	Gag-Pol (p2, p7, p1, p6)
1001-1100	8	Gag-Pol (Protease)
1101-1200	6	Gag-Pol (RT)
1201-1300	8	Gag-Pol (RT)
1301-1400	9	Gag-Pol (RT)
1401-1500	6	Gag-Pol (RT)
1501-1600	14	Gag-Pol (RT)
1601-1700	6	Gag-Pol (RT, Integrase)
1701-1800	15	Gag-Pol (Integrase)
1801-1900	4	Gag-Pol (Integrase)
1901-2000	16	Gag-Pol (Integrase), Env gp160
2001-2100	13	Env gp160
2101-2200	9	Env gp160

Continued...

Table G.1 – continued from previous page		
Position	N	**HIV-1 protein**
2201-2300	10	Env gp160
2301-2400	14	Env gp160
2401-2500	3	Env gp160
2501-2600	10	Env gp160
2601-2700	10	Env gp160
2701-2800	6	Env gp160
2801-2900	3	Nef
2901-3000	15	Nef
3001-3100	10	Tat, Vpu
3101-3200	6	Vpu, Rev
3201-3300	12	Rev, Vif
3301-3400	10	Vif
3401-3500	12	Vif, Vpr
3501-3535	3	Vpr

The HIV-1 proteome is broken up into separate proteins at the 5-mer level.

Table G.2: Table of counts of non-overlapping 5-mers by dividing the HIV-1 proteome into individual protein segments. N is the observed number of 5-mers starting within the protein with no overlap in the human proteome. Expected N is calculated as (total # of non-overlaps)$\times$(# of 5-mer from this protein)$\div$(total # of 5-mers) = 336$\times$(# of 5-mer from this protein)$\div$3535.

Protein	**Position**	N	**Expected** N
Gag poly (p17)	0001-0131	4	12.451
Gag poly (p24)	0132-0362	26	21.956
Gag poly (p2, p7, p1, p6)	0363-0507	18	13.782
Gag-Pol (p17)	0508-0638	4	12.451
Gag-Pol (p24)	0639-0869	26	21.956
Gag-Pol (p2, p7, p1, p6)	0870-1006	21	13.022
Gag-Pol (Protease)	1007-1105	8	9.410
Gag-Pol (RT)	1106-1665	44	53.228
			Continued...

Table G.2 – continued from previous page			
Protein	**Position**	N	**Expected** N
Gag-Pol (Integrase)	1666-1949	25	26.994
Env gp160	1950-2801	89	80.982
Nef	2802-3002	18	19.105
Tat	3003-3084	10	7.794
Vpu	3085-3161	6	7.319
Rev	3162-3273	1	10.646
Vif	3274-3461	27	17.869
Vpr	3462-3535	9	7.034

The HIV-1 proteome is broken up into 36 segments of 100 amino acids in length at the 6-mer level.

Table G.3: Table of counts of overlapping 6-mers by dividing the HIV-1 proteome into evenly-sized (i.e., 100 amino acids) segments. N is the observed number of 6-mers in the segment which overlap the human proteome.

Position	N	**HIV-1 protein**
0001-0100	46	Gag poly (p17)
0101-0200	32	Gag poly (p17, p24)
0201-0300	29	Gag poly (p24)
0301-0400	25	Gag poly (p24, p2, p7, p1, p6)
0401-0500	38	Gag poly (p2, p7, p1, p6)
0501-0600	42	Gag-Pol (p17)
0601-0700	37	Gag-Pol (p17, p24)
0701-0800	28	Gag-Pol (p24)
0801-0900	25	Gag-Pol (p24, p2, p7, p1, p6)
0901-1000	30	Gag-Pol (p2, p7, p1, p6)
1001-1100	26	Gag-Pol (Protease)
1101-1200	22	Gag-Pol (RT)
1201-1300	25	Gag-Pol (RT)
1301-1400	33	Gag-Pol (RT)
1401-1500	30	Gag-Pol (RT))
1501-1600	21	Gag-Pol (RT)
		Continued...

Table G.3 – continued from previous page		
Position	N	**HIV-1 protein**
1601-1700	20	Gag-Pol (RT, Integrase)
1701-1800	22	Gag-Pol (Integrase)
1801-1900	20	Gag-Pol (Integrase)
1901-2000	21	Gag-Pol (Integrase), Env gp160
2001-2100	15	Env gp160
2101-2200	18	Env gp160
2201-2300	28	Env gp160
2301-2400	17	Env gp160
2401-2500	31	Env gp160
2501-2600	34	Env gp160
2601-2700	46	Env gp160
2701-2800	41	Env gp160
2801-2900	42	Nef
2901-3000	24	Nef
3001-3100	31	Tat, Vpu
3101-3200	48	Vpu, Rev
3201-3300	32	Rev, Vif
3301-3400	28	Vif
3401-3500	34	Vif, Vpr
3501-3526	8	Vpr

The HIV-1 proteome is broken up into separate proteins at the 6-mer level.

Table G.4: Table of counts of overlapping 6-mers by dividing the HIV-1 proteome into individual protein segments. N is the observed number of 6-mers starting within the protein with overlap in the human proteome. Expected N is calculated as (total # of overlaps)$\times$(# of 6-mer from this protein)$\div$(total # of 6-mers) = 1049$\times$(# of 6-mer from this protein)$\div$3526.

Protein	**Position**	N	**Expected** N
Gag poly (p17)	0001-0131	58	38.973
Gag poly (p24)	0132-0362	68	68.723
			Continued...

Table G.4 – continued from previous page			
Protein	**Position**	N	**Expected** N
Gag poly (p2, p7, p1, p6)	0363-0506	45	42.841
Gag-Pol (p17)	0507-0637	58	38.973
Gag-Pol (p24)	0638-0868	68	68.723
Gag-Pol (p2, p7, p1, p6)	0869-1005	35	40.758
Gag-Pol (Protease)	1006-1104	27	29.453
Gag-Pol (RT)	1105-1664	146	166.602
Gag-Pol (Integrase)	1665-1947	59	84.194
Env gp160	1948-2798	237	253.176
Nef	2799-2998	66	59.501
Tat	2999-3079	23	24.098
Vpu	3080-3155	32	22.61
Rev	3156-3266	54	33.023
Vif	3267-3453	53	55.633
Vpr	3454-3526	20	21.718

The HIV-1 proteome is broken up into 18 segments of 200 amino acids in length at the 7-mer level.

Table G.5: Table of counts of overlapping 7-mers by dividing the HIV-1 proteome into evenly-sized (i.e., 200 amino acids) segments. N is the observed number of 7-mers in the segment which overlap the human proteome.

Position	N	**HIV-1 protein**
0001-0200	7	Gag poly (p17)
0201-0400	8	Gag poly (p17, p24)
0401-0600	9	Gag poly (p2, p7, p1, p6), Gag-Pol (p17)
0601-0800	5	Gag-Pol (p17, p24)
0801-1000	7	Gag-Pol (p24, p2, p7, p1, p6)
1001-1200	3	Gag-Pol (Protease, RT)
1201-1400	7	Gag-Pol (RT)
1401-1600	8	Gag-Pol (RT)
1601-1800	1	Gag-Pol (RT, Integrase)
1801-2000	3	Gag-Pol (Integrase), Env gp160
Continued...		

Table G.5 – continued from previous page		
Position	N	**HIV-1 protein**
2001-2200	3	Env gp160
2201-2400	4	Env gp160
2401-2600	6	Env gp160
2601-2800	17	Env gp160
2801-3000	5	Nef
3001-3200	10	Tat, Vpu, Rev
3201-3400	5	Rev, Vif
3401-3517	4	Vif, Vpr

The HIV-1 proteome is broken up into separate proteins at the 7-mer level.

Table G.6: Table of counts of overlapping 7-mers by dividing the HIV-1 proteome into individual protein segments. N is the observed number of 7-mers starting within the protein with overlap in the human proteome. Expected N is calculated as (total # of overlaps)$\times$(# of 7-mer from this protein)$\div$(total # of 7-mers) = 112$\times$(# of 7-mer from this protein)$\div$3517.

Protein	**Position**	N	**Expected** N
Gag poly (p17)	0001-0131	7	4.172
Gag poly (p24)	0132-0362	7	7.356
Gag poly (p2, p7, p1, p6)	0363-0505	6	4.554
Gag-Pol (p17)	0506-0636	7	4.172
Gag-Pol (p24)	0637-0867	7	7.356
Gag-Pol (p2, p7, p1, p6)	0868-1004	2	4.363
Gag-Pol (Protease)	1005-1103	3	3.153
Gag-Pol (RT)	1104-1663	15	17.833
Gag-Pol (Integrase)	1664-1945	3	8.98
Env gp160	1946-2795	31	27.069
Nef	2796-2994	5	6.337
Tat	2995-3074	3	2.548
Vpu	3075-3149	4	2.388
Rev	3150-3259	6	3.503
			Continued...

Table G.6 – continued from previous page			
Protein	**Position**	N	**Expected** N
Vif	3260-3445	5	5.923
Vpr	3446-3517	1	2.293

The HIV-2 proteome is broken up into 37 segments of 100 amino acids in length at the 5-mer level.

Table G.7: Table of counts of non-overlapping 5-mers by dividing the HIV-2 proteome into evenly-sized (i.e., 100 amino acids) segments. N is the observed number of 5-mers in the segment with no overlap in the human proteome.

Position	N	**HIV-2 protein**
0001-0100	9	Env gp160
0101-0200	8	Env gp160
0201-0300	21	Env gp160
0301-0400	25	Env gp160
0401-0500	12	Env gp160
0501-0600	2	Env gp160
0601-0700	13	Env gp160
0701-0800	10	Env gp160
0801-0900	6	Env gp160, Gag-Pol (p17)
0901-1000	9	Gag-Pol (p17, p24)
1001-1100	11	Gag-Pol (p24)
1101-1200	15	Gag-Pol (p24)
1201-1300	11	Gag-Pol (p24, p2, p7, p1, p6)
1301-1400	4	Gag-Pol (p2, p7, p1, p6, Protease)
1401-1500	5	Gag-Pol (Protease, RT)
1501-1600	6	Gag-Pol (RT)
1601-1700	11	Gag-Pol (RT)
1701-1800	8	Gag-Pol (RT)
1801-1900	15	Gag-Pol (RT)
1901-2000	3	Gag-Pol (RT)
2001-2100	14	Gag-Pol (RT, Integrase)
2101-2200	12	Gag-Pol (Integrase)
		Continued...

Table G.7 – continued from previous page		
Position	N	**HIV-2 protein**
2201-2300	7	Gag-Pol (Integrase)
2301-2400	7	Gag-Pol (Integrase), Gag poly (p17)
2401-2500	8	Gag poly (p17, p24)
2501-2600	12	Gag poly (p24)
2601-2700	10	Gag poly (p24, p2, p7, p1, p6)
2701-2800	12	Gag poly (p2, p7, p1, p6)
2801-2900	15	Gag poly (p2, p7, p1, p6), Vif
2901-3000	17	Vif
3001-3100	4	Vif, Nef
3101-3200	16	Nef
3201-3300	11	Nef, Tat
3301-3400	11	Tat
3401-3500	5	Tat, Rev
3501-3600	4	Rev, Vpr
3601-3700	13	Vpr, Vpx

The HIV-2 proteome is broken up into separate proteins at the 5-mer level.

Table G.8: Table of counts of non-overlapping 5-mers by dividing the HIV-2 proteome into individual protein segments. N is the observed number of 5-mers starting within the protein with no overlap in the human proteome. Expected N is calculated as (total # of non-overlaps)$\times$(# of 5-mer from this protein)$\div$(total # of 5-mers) = 382$\times$(# of 5-mer from this protein)$\div$3723.

Protein	**Position**	N	**Expected** N
Env gp160	0001-0854	105	87.625
Gag-Pol (p17)	0855-0988	9	13.749
Gag-Pol (p24)	0989-1218	28	23.599
Gag-Pol (p2,p7,p1,p6)	1219-1366	14	15.186
Gag-Pol (Protease)	1367-1465	2	10.158
Gag-Pol (RT)	1466-2024	47	57.356
Gag-Pol (Integrase)	2025-2313	32	29.653
Continued...			

Table G.8 – continued from previous page			
Protein	**Position**	N	**Expected** N
Gag poly (p17)	2314-2447	9	13.749
Gag poly (p24)	2448-2677	28	23.599
Gag poly (p2,p7,p1,p6)	2678-2830	12	15.699
Vif	2831-3041	33	21.65
Nef	3042-3292	30	25.754
Tat	3293-3418	11	12.928
Rev	3419-3514	5	9.85
Vpr	3515-3615	5	10.363
Vpx	3616-3723	12	11.081

The HIV-2 proteome is broken up into 37 segments of 100 amino acids in length at the 6-mer level.

Table G.9: Table of counts of overlapping 6-mers by dividing the HIV-2 proteome into evenly-sized (i.e., 100 amino acids) segments. N is the observed number of 6-mers in the segment which overlap the human proteome.

Position	N	**HIV-2 protein**
0001-0100	22	Env gp160
0101-0200	36	Env gp160
0201-0300	11	Env gp160
0301-0400	16	Env gp160
0401-0500	22	Env gp160
0501-0600	48	Env gp160
0601-0700	28	Env gp160
0701-0800	31	Env gp160
0801-0900	41	Env gp160, Gag-Pol (p17)
0901-1000	38	Gag-Pol (p17, p24)
1001-1100	29	Gag-Pol (p24)
1101-1200	12	Gag-Pol (p24)
1201-1300	45	Gag-Pol (p24, p2, p7, p1, p6)
1301-1400	42	Gag-Pol (p2, p7, p1, p6, Protease)
1401-1500	32	Gag-Pol (Protease, RT)
Continued...		

Table G.9 – continued from previous page		
Position	N	**HIV-2 protein**
1501-1600	35	Gag-Pol (RT)
1601-1700	27	Gag-Pol (RT)
1701-1800	30	Gag-Pol (RT)
1801-1900	23	Gag-Pol (RT)
1901-2000	34	Gag-Pol (RT)
2001-2100	16	Gag-Pol (RT, Integrase)
2101-2200	26	Gag-Pol (Integrase)
2201-2300	24	Gag-Pol (Integrase)
2301-2400	44	Gag-Pol (Integrase), Gag poly (p17)
2401-2500	31	Gag poly (p17, p24)
2501-2600	22	Gag poly (p24)
2601-2700	29	Gag poly (p24, p2, p7, p1, p6)
2701-2800	36	Gag poly (p2, p7, p1, p6)
2801-2900	17	Gag poly (p2, p7, p1, p6), Vif
2901-3000	27	Vif
3001-3100	46	Vif, Nef
3101-3200	31	Nef
3201-3300	26	Nef, Tat
3301-3400	41	Tat
3401-3500	34	Tat, Rev
3501-3600	44	Rev, Vpr
3601-3700	30	Vpr, Vpx
3701-3714	12	Vpx

The HIV-2 proteome is broken up into separate proteins at the 6-mer level.

Table G.10: Table of counts of overlapping 6-mers by dividing the HIV-2 proteome into individual protein segments. N is the observed number of 6-mers starting within the protein with overlap in the human proteome. Expected N is calculated as (total # of overlaps)$\times$(# of 6-mer from this protein)$\div$(total # of 6-mers) = 1138$\times$(# of 6-mer from this protein)$\div$3714.

Protein	**Position**	N	**Expected** N
Env gp160	0001-0853	235	261.366
Gag-Pol (p17)	0854-0987	55	41.059
Gag-Pol (p24)	0988-1217	55	70.474
Gag-Pol (p2,p7,p1,p6)	1218-1365	64	45.348
Gag-Pol (Protease)	1366-1464	29	30.334
Gag-Pol (RT)	1465-2023	170	171.282
Gag-Pol (Integrase)	2024-2311	66	88.246
Gag poly (p17)	2312-2445	55	41.059
Gag poly (p24)	2446-2675	55	70.474
Gag poly (p2,p7,p1,p6)	2676-2827	53	46.574
Vif	2828-3037	58	64.346
Nef	3038-3287	76	76.602
Tat	3288-3412	50	38.301
Rev	3413-3507	35	29.109
Vpr	3508-3607	43	30.641
Vpx	3608-3714	39 32.786	

The HIV-2 proteome is broken up into 25 segments of 150 amino acids in length at the 7-mer level.

Table G.11: Table of counts of overlapping 7-mers by dividing the HIV-2 proteome into evenly-sized (i.e., 150 amino acids) segments. N is the observed number of 7-mers in the segment which overlap the human proteome.

Position	N	**HIV-2 protein**
0001-0150	6	Env gp160
0151-0300	1	Env gp160
		Continued...

Table G.11 – continued from previous page		
Position	N	**HIV-2 protein**
0301-0450	2	Env gp160
0451-0600	11	Env gp160
0601-0750	0	Env gp160
0751-0900	11	Env gp160, Gag-Pol (p17)
0901-1050	5	Gag-Pol (p17, p24)
1051-1200	3	Gag-Pol (p24)
1201-1350	17	Gag-Pol (p24, p2, p7, p1, p6)
1351-1500	5	Gag-Pol (p2, p7, p1, p6, Protease, RT)
1501-1650	2	Gag-Pol (RT)
1651-1800	4	Gag-Pol (RT)
1801-1950	5	Gag-Pol (RT)
1951-2100	2	Gag-Pol (RT, Integrase)
2101-2250	2	Gag-Pol (Integrase)
2251-2400	7	Gag-Pol (Integrase), Gag poly (p17)
2401-2550	5	Gag poly (p17, p24)
2551-2700	7	Gag poly (p24, p2, p7, p1, p6)
2701-2850	8	Gag poly (p2, p7, p1, p6), Vif
2851-3000	2	Vif
3001-3150	6	Vif, Nef
3151-3300	3	Nef, Tat
3301-3450	7	Tat, Rev
3451-3600	6	Rev, Vpr
3601-3705	11	Vpx

The HIV-2 proteome is broken up into separate proteins at the 7-mer level.

Table G.12: Table of counts of overlapping 7-mers by dividing the HIV-2 proteome into individual protein segments. N is the observed number of 7-mers starting within the protein with overlap in the human proteome. Expected N is calculated as (total # of overlaps)$\times$(# of 7-mer from this protein)$\div$(total # of 7-mers) = 138$\times$(# of 7-mer from this protein)$\div$3705.

Protein	**Position**	N	**Expected** N
Env gp160	0001-0852	27	31.734
Gag-Pol (p17)	0853-0986	8	4.991
Gag-Pol (p24)	0987-1216	6	8.567
Gag-Pol (p2,p7,p1,p6)	1217-1364	16	5.513
Gag-Pol (Protease)	1365-1463	2	3.687
Gag-Pol (RT)	1464-2022	14	20.821
Gag-Pol (Integrase)	2023-2309	5	10.690
Gag poly (p17)	2310-2443	8	4.991
Gag poly (p24)	2444-2673	6	8.567
Gag poly (p2,p7,p1,p6)	2674-2824	11	5.624
Vif	2825-3033	3	7.785
Nef	3034-3282	7	9.274
Tat	3283-3406	3	4.619
Rev	3407-3500	5	3.501
Vpr	3501-3599	6	3.687
Vpx	3600-3705	11	3.948

The Influenza A virus proteome is broken up into 45 segments of 100 amino acids in length at the 5-mer level.

Table G.13: Table of counts of non-overlapping 5-mers by dividing the Influenza A virus proteome into evenly-sized (i.e., 100 amino acids) segments. N is the observed number of 5-mers in the segment with no overlap in the human proteome.

Position	N	**Influenza A virus protein**
0001-0100	13	Nonstructural 1
		Continued...

Table G.13 – continued from previous page		
Position	N	**Influenza A virus protein**
0101-0200	6	Nonstructural 1
0201-0300	8	Nonstructural 1, Nonstructural 2
0301-0400	3	Nonstructural 2, Matrix 1
0401-0500	2	Matrix 1
0501-0600	6	Matrix 1, Matrix 2
0601-0700	11	Matrix 2, Hemagglutinin
0701-0800	9	Hemagglutinin
0801-0900	18	Hemagglutinin
0901-1000	12	Hemagglutinin
1001-1100	7	Hemagglutinin
1101-1200	7	Hemagglutinin
1201-1300	8	Hemagglutinin, Neuraminidase
1301-1400	8	Neuraminidase
1401-1500	9	Neuraminidase
1501-1600	15	Neuraminidase
1601-1700	20	Neuraminidase
1701-1800	7	Neuraminidase, Nucleocapsid
1801-1900	12	Nucleocapsid
1901-2000	10	Nucleocapsid
2001-2100	8	Nucleocapsid
2101-2200	6	Nucleocapsid
2201-2300	8	Nucleocapsid, Polymerse (gene: None)
2301-2400	2	Polymerse (gene: None)
2401-2500	8	Polymerse (gene: None)
2501-2600	8	Polymerse (gene: None)
2601-2700	10	Polymerse (gene: None)
2701-2800	8	Polymerse (gene: None)
2801-2900	5	Polymerse (gene: None)
2901-3000	10	Polymerse (gene: None), Polymerse (gene: PB1)
3001-3100	8	Polymerse (gene: PB1)
3101-3200	6	Polymerse (gene: PB1)
3201-3300	14	Polymerse (gene: PB1)
3301-3400	15	Polymerse (gene: PB1)
Continued...		

Table G.13 – continued from previous page		
Position	N	**Influenza A virus protein**
3401-3500	11	Polymerse (gene: PB1)
3501-3600	7	Polymerse (gene: PB1)
3601-3700	3	Polymerse (gene: PB1), Polymerse (gene: PB2)
3701-3800	11	Polymerse (gene: PB2)
3801-3900	4	Polymerse (gene: PB2)
3901-4000	7	Polymerse (gene: PB2)
4001-4100	4	Polymerse (gene: PB2)
4101-4200	8	Polymerse (gene: PB2)
4201-4300	13	Polymerse (gene: PB2)
4301-4400	0	Polymerse (gene: PB2)
4401-4427	1	Polymerse (gene: PB2)

The Influenza A virus proteome is broken up into separate proteins at the 5-mer level.

Table G.14: Table of counts of non-overlapping 5-mers by dividing the Influenza A virus proteome into individual protein segments. N is the observed number of 5-mers starting within the protein with no overlap in the human proteome. Expected N is calculated as (total # of non-overlaps)×(# of 5-mer from this protein)÷(total # of 5-mers) = 376×(# of 5-mer from this protein)÷4427.

Protein	**Position**	N	**Expected** N
Nonstructural 1	0001-0226	21	19.195
Nonstructural 2	0227-0343	8	9.937
Matrix 1	0344-0591	9	21.063
Matrix 2	0592-0684	10	7.900
Hemagglutinin	0685-1248	58	47.902
Neuraminidase	1249-1713	58	39.494
Nucleocapsid	1714-2207	41	41.957
Polymerse (gene: None)	2208-2919	53	60.473
Polymerse (gene: PB1)	2920-3672	68	63.955
Polymerse (gene: PB2)	3673-4427	50	64.125

The Influenza A virus proteome is broken up into 45 segments of 100 amino acids in length at the 6-mer level.

Table G.15: Table of counts of overlapping 6-mers by dividing the Influenza A virus proteome into evenly-sized (i.e., 100 amino acids) segments. N is the observed number of 6-mers in the segment which overlap the human proteome.

Position	N	**Influenza A virus protein**
0001-0100	34	Nonstructural 1
0101-0200	33	Nonstructural 1
0201-0300	28	Nonstructural 1, Nonstructural 2
0301-0400	35	Nonstructural 2, Matrix 1
0401-0500	32	Matrix 1
0501-0600	33	Matrix 1, Matrix 2
0601-0700	32	Matrix 2, Hemagglutinin
0701-0800	25	Hemagglutinin
0801-0900	21	Hemagglutinin
0901-1000	14	Hemagglutinin
1001-1100	31	Hemagglutinin
1101-1200	27	Hemagglutinin
1201-1300	31	Hemagglutinin, Neuraminidase
1301-1400	28	Neuraminidase
1401-1500	28	Neuraminidase
1501-1600	14	Neuraminidase
1601-1700	22	Neuraminidase
1701-1800	32	Neuraminidase, Nucleocapsid
1801-1900	22	Nucleocapsid
1901-2000	25	Nucleocapsid
2001-2100	28	Nucleocapsid
2101-2200	22	Nucleocapsid
2201-2300	18	Polymerse (gene: None)
2301-2400	30	Polymerse (gene: None)
2401-2500	31	Polymerse (gene: None)
2501-2600	28	Polymerse (gene: None)
2601-2700	22	Polymerse (gene: None)
2701-2800	24	Polymerse (gene: None)
		Continued...

Table G.15 – continued from previous page		
Position	N	**Influenza A virus protein**
2801-2900	39	Polymerse (gene: None)
2901-3000	22	Polymerse (gene: None), Polymerse (gene: PB1)
3001-3100	23	Polymerse (gene: PB1)
3101-3200	34	Polymerse (gene: PB1)
3201-3300	16	Polymerse (gene: PB1)
3301-3400	22	Polymerse (gene: PB1)
3401-3500	25	Polymerse (gene: PB1)
3501-3600	18	Polymerse (gene: PB1)
3601-3700	27	Polymerse (gene: PB1), Polymerse (gene: PB2)
3701-3800	13	Polymerse (gene: PB2)
3801-3900	34	Polymerse (gene: PB2)
3901-4000	38	Polymerse (gene: PB2)
4001-4100	29	Polymerse (gene: PB2)
4101-4200	25	Polymerse (gene: PB2)
4201-4300	34	Polymerse (gene: PB2)
4301-4400	40	Polymerse (gene: PB2)
4401-4417	6	Polymerse (gene: PB2)

The Influenza A virus proteome is broken up into separate proteins at the 6-mer level.

Table G.16: Table of counts of overlapping 6-mers by dividing the Influenza A virus proteome into individual protein segments. N is the observed number of 6-mers starting within the protein with overlap in the human proteome. Expected N is calculated as (total # of overlaps)$\times$(# of 6-mer from this protein)$\div$(total # of 6-mers) = 1195$\times$(# of 6-mer from this protein)$\div$4417.

Protein	**Position**	N	**Expected** N
Nonstructural 1	0001-0225	75	60.873
Nonstructural 2	0226-0341	33	31.383
Matrix 1	0342-0588	85	66.825
Matrix 2	0589-0680	23	24.890
Hemagglutinin	0681-1243	144	152.317
Continued...			

Table G.16 – continued from previous page			
Protein	**Position**	N	**Expected** N
Neuraminidase	1244-1707	111	125.533
Nucleocapsid	1708-2200	126	133.379
Polymerse (gene: None)	2201-2911	195	192.358
Polymerse (gene: PB1)	2912-3663	172	203.450
Polymerse (gene: PB2)	3664-4417	231	203.991

The Influenza A virus proteome is broken up into 18 segments of 250 amino acids in length at the 7-mer level.

Table G.17: Table of counts of overlapping 7-mers by dividing the Influenza A virus proteome into evenly-sized (i.e., 250 amino acids) segments. N is the observed number of 7-mers in the segment which overlap the human proteome.

Position	N	**Influenza A virus protein**
0001-0250	12	Nonstructural 1, Nonstrucrual 2
0251-0500	5	Nonstrucrual 2, Matrix 1
0501-0750	2	Matrix 1, Matrix 2, Hemagglutinin
0751-1000	4	Hemagglutinin
1001-1250	9	Hemagglutinin, Neuraminidase
1251-1500	4	Neuraminidase
1501-1750	4	Neuraminidase, Nucleocapsid
1751-2000	5	Nucleocapsid
2001-2250	11	Nucleocapsid, Polymerse (gene: None)
2251-2500	8	Polymerse (gene: None)
2501-2750	3	Polymerse (gene: None)
2751-3000	6	Polymerse (gene: None), Polymerse (gene: PB1)
3001-3250	3	Polymerse (gene: PB1)
3251-3500	2	Polymerse (gene: PB1)
3501-3750	0	Polymerse (gene: PB1), Polymerse (gene: PB2)
3751-4000	8	Polymerse (gene: PB2)
4001-4250	5	Polymerse (gene: PB2)
4251-4407	5	Polymerse (gene: PB2)

The Influenza A virus proteome is broken up into separate proteins at the 7-mer level.

Table G.18: Table of counts of overlapping 7-mers by dividing the Influenza A virus proteome into individual protein segments. N is the observed number of 7-mers starting within the protein with overlap in the human proteome. Expected N is calculated as (total # of overlaps)$\times$(# of 7-mer from this protein)$\div$(total # of 7-mers) = 96$\times$(# of 7-mer from this protein)$\div$4407.

Protein	**Position**	N	**Expected** N
Nonstructural 1	0001-0224	10	4.880
Nonstructural 2	0225-0339	4	2.505
Matrix 1	0340-0585	4	5.359
Matrix 2	0586-0676	1	1.982
Hemagglutinin	0677-1238	12	12.242
Neuraminidase	1239-1701	7	10.086
Nucleocapsid	1702-2193	16	10.717
Polymerse (gene: None)	2194-2903	19	15.466
Polymerse (gene: PB1)	2904-3654	5	16.359
Polymerse (gene: PB2)	3655-4407	18	16.403